EAGLES

Expert Advisory Groups on Language Engineering Standards

Spoken Language Working Group EC-DGXIII, LRE LRE-61-100

HANDBOOK

of

Standards and Resources for Spoken Language Systems

Edited by

Dafydd Gibbon, Roger Moore, Richard Winski

Editorial preface

The technical production of this handbook has been a joint effort by several groups involving a large number of people, and the success of the coordination process in itself is by no means the least significant result of the EAGLES Spoken Language Working Group (see Chapter 1), with particular credit to the fellow-members of the editorial team, Roger Moore and Richard Winski, for an inspiring and supportive style of collaboration.

The first group comprises the technical authors. All of them deserve thanks for their patience and cooperativeness, despite over-full schedules, heavy responsibilities and, in many cases, also the need to learn IAT_EX in the process.

The second group includes the EAGLES support team, particularly the organisers in Pisa, Antonio Zampolli and Nicoletta Calzolari, with their untiring efforts to coordinate a somewhat unruly band of experts. Jock McNaught brought his editorial expertise in electronic publishing to bear on the initial layout design, and on solutions for a multitude of thorny problems.

The third group includes my team in Bielefeld, especially Inge Mertins, who put in a massive amount of work researching sources, re-formatting from a variety of source formats, taking care of complex style packages and spending month after month gently and effectively coaxing various authors to provide readable text, graphics and formulae. Thorsten Bomberg brought his expert knowledge of UNIX systems programming to bear on many technical problems; having found that available software did not scale up to handle a document of the size and complexity of this handbook, he specified and implemented a LATEX to HTML conversion strategy which did work, and shared his results with the latex2html software developers, resulting in better software. Holger Ulrich Nord and Thorsten Trippel re-formatted the revised version in HTML, battling with many new format styles.

The fourth group is the publishing team led by Anke Beck at Mouton de Gruyter, whose professional standards forced us all to re-think many aspects of presentation and formatting, and on whose advice we were able to rely in designing a LATEX document class to emulate the Mouton de Gruyter house style (though a couple of our own oddities remain).

The fifth group comprises those responsible at Directorate General XIII of the European Commission, Norbert Brinkhoff-Button, the project officer, and Roberto Cencioni, who deserve acknowledgment for their foresight, their willingness to be persuaded to take risks with this novel publishing venture for the field of spoken language technology, and above all for their patience in what must have seemed like an unending production story.

The main aim during the technical production process was to produce a high-quality handbook which on the one hand documents the core of stan-

vi Editorial preface

dard good practice during the 1990s, and on the other hand presents a solid platform for further development. To attain this goal, a number of textual smoothing processes were required. The format conversion and formatting tasks have already been mentioned; English style and idiom in several chapters, by both native and non-native speakers, had to be considerably adapted for general readability and consistency. Many overlaps were removed, many additional details incorporated, cross-references to other chapters and the other EAGLES Working Groups were included, copyrights (for instance for electronic IPA versions) were negotiated, and additional appendix materials were elicited. Some of the appendices were specially written for the handbook, but most were generously provided by other European Commission funded projects, notably the SAM project, and were left unchanged apart from the necessary re-formatting. In certain areas, for instance, with corpus copyrights and with clandestine recording, legal and ethical issues arose, which could only be touched on in passing.

Recommendations are given explicitly in subsections in each chapter, and can thus be conveniently referred to by consulting the table of contents, which is deliberately kept rather detailed and is thus unusually long. The task of completely 'homogenising' the style of recommendations proved to be too comprehensive at the present stage, however, partly because of the variety of recommendation types, and partly because of the different presentation styles of authors from different disciplines.

Since the original conception of the report four years before publication, the importance of the World Wide Web for research has expanded enormously. This has made the publication of sources for corpora and tools unnecessary: Web search engines can quickly find the up-to-date addresses. The second consideration which emerged shortly before the final production phase was the possibility of publication on the Web. The pros and cons of this were much debated, and criteria of overall portability, durability, robustness and convenience of paper versions (with library and paperback editions) scored over a purely electronic hypertext mode; in addition, the publisher is providing CD-ROM and, courageously, Web versions.

Despite all efforts, the handbook has a number of obvious shortcomings, and readers will no doubt collect their own selection of these. For the shortcomings I beg the readers' indulgence, and urge them to communicate their suggestions and thereby help to improve future versions of the handbook.

> Dafydd Gibbon (Technical Editor)

Main technical authors

User's guide

Roger Moore

Part I:	
System design	Khalid Choukri
SL corpus design	Els den Os
SL corpus collection	Christoph Draxler
SL corpus representation	Els den Os

Part II:

SL lexicaDafydd GibbonLanguage modelsHermann NeyPysical characterisation and
descriptionLars Knohl & Volker Kraft

Part III:

Assessment methodologies and	Peter Howell
experimental design	
Assessment of recognition systems	David van Leeuwen & H. Steeneken
Assessment of speaker verification	Frédéric Bimbot &
systems	Gérard Chollet
Assessment of synthesis systems	Renée van Bezooijen &
	Vincent van Heuven
Assessment of interactive systems	Norman Fraser

Part IV:

Contributions from named ProjectsMain new contributionsCand revisionsJo

Christoph Draxler, John Esling, John Wells

viii Technical authors

Contents

1	User	's guid	e	1
	1.1	Backg	round	1
		1.1.1	EAGLES objectives	1
		1.1.2	EAGLES organisational structure	2
		1.1.3	EAGLES workplan	3
	1.2	Spoke	n Language systems, standards and resources	4
		1.2.1	Spoken Language systems	4
		1.2.2	Standards and resources for Spoken Language systems	5
	1.3	The E	AGLES Spoken Language Working Group (WG5)	9
		1.3.1	Subgroups of the EAGLES Spoken Language Working	
			Group	11
		1.3.2	Relationships with the other EAGLES Working Groups	11
		1.3.3	Workshops	11
		1.3.4	Production of the handbook	13
		1.3.5	Consultation with the R&D Community	13
	1.4	Overv	iew of the handbook	14
		1.4.1	Intended readership	14
		1.4.2	Scope	15
		1.4.3	The main chapters of the handbook	21
	1.5	The c	urrent state of play	25
	1.6	Possib	ble future actions	25
		1.6.1	Revision and completion of existing documentation .	25
		1.6.2	Extended survey of existing practice	25
		1.6.3	Extension of language base	26
		1.6.4	Terminology	26
		1.6.5	Move to prescriptive recommendations	26
		1.6.6	Publication and dissemination	26
		1.6.7	Coordination with other bodies	26
	1.7	Conta	ct points	27
	1.8	Ackno	wledgements	28
D,	vrt I.	Sn	okan languaga system and corpus design	20
1 6	art r.	ъþ	oken language system and corpus design	29
2	Syst	em des	ign	30
	2.1	Introd	luction	30
		2.1.1	System capability profile versus application require-	
			ment profile	32
		2.1.2	Technical features versus system capabilities	33
		2.1.3	System in operation versus laboratory prototype \ldots	34

2.2	Speech	1 recognition systems	34
	2.2.1	Speaker dependency	35
	2.2.2	Speaker-adapted systems	36
	2.2.3	Speaking aspects	37
	2.2.4	Vocabulary aspects	38
	2.2.5	Branching factor (perplexity factor)	41
	2.2.6	Rejection mode	41
	2.2.7	Application vocabulary and confusion matrix	42
	2.2.8	Speaker selection	43
	2.2.9	Language modelling	43
	2.2.10	Channel adaptation/Environment adaptation	43
	2.2.11	Task/Application adaptation	44
	2.2.12	Speech recognition and application interfaces	44
	2.2.13	Speech input and speech signal acquisition	44
	2.2.14	Cut-through versus voice-stop (anticipation)	48
	2.2.15	Error measure presentation	49
	2.2.16	Error handling	49
	2.2.17	Response time	50
2.3	Speake	er verification/identification	51
	2.3.1	Speaker verification versus speaker identification \ldots	51
	2.3.2	Speaker characteristics verification or identification	51
	2.3.3	Error measure	51
	2.3.4	Training	53
	2.3.5	Exploitation	54
2.4	Speech	1 synthesis	56
	2.4.1	Speech recording, storage, and playback \ldots .	56
	2.4.2	$Canned speech \dots \dots$	58
	2.4.3	Text-to-speech synthesis	58
	2.4.4	Multi-linguality	61
2.5	Intera	ctive voice systems	61
	2.5.1	Menu-driven systems	61
	2.5.2	Sophisticated dialogue manager	62
	2.5.3	Interaction with an information database	62
	2.5.4	Error recovery strategy	62
	2.5.5	Routing to an operator	63
	2.5.6	DTMF, pulse detection, and voice recognition	63
2.6	System	n platform	63
	2.6.1	Software aspects	63
	2.6.2	Hardware aspects	64
a =	2.6.3	Planning for expansion	67
2.7	Systen	n simulation and prototyping	67
2.8	Host-c	computer interfaces	67
2.9	Comp	uter telephony integration	67

$\operatorname{Contents}$	xi

	$2.10 \\ 2.11 \\ 2.12$	Multi- System Statist	lingual aspects	68 68 70
	$\begin{array}{c} 2.13\\ 2.14\end{array}$	Conclu Recom	nsion	70 70
3	SL c	orpus d	lesign	79
	3.1	Introd	uction	79
		3.1.1	Spoken language corpus	79
		3.1.2	About this chapter	80
	3.2	Eight 1	main differences between collections of written and spo-	
		ken lai	nguage data	81
		3.2.1	Durability of text, volatility of speech	82
		3.2.2	Different production times for text and speech	82
		3.2.3	Correcting errors in the production of text and speech	82
		3.2.4	orthographic identity and phonetic variability of lex-	02
		325	Printable ASCII-strings and continuously sampled	00
		0.2.0	speech	84
		3.2.6	Size differences between NL and SL data	85
		3.2.7	The different legal status of written texts and spoken	
			words	85
		3.2.8	The different nature of categories and time functions .	86
	3.3	Applic	ations of spoken language corpora	86
		3.3.1	Speech corpora for research purposes	87
		3.3.2	Speech corpora for technological applications	92
	3.4	Specifi	cation of the linguistic content	99
		3.4.1	Different types of speech data	99
	0 5	3.4.2	Factorial experiments and corpus studies	106
	3.5	Specin	Carpus size in terms of speakers	107
		3.3.1 3.5.9	Speaker characteristics	1107
	3.6	5.5.2 Final o	comments	118
	0.0	I IIIGI (110
4	SL c	orpus c	ollection	119
	4.1	Introd	uction	119
	4.2	Data c	collection dimensions	120
		4.2.1	Visibility: open vs. secret	120
		4.2.2	Clandestine microphone recordings	120
		4.2.3	Visible recordings	121
		4.2.4	Environment: Studio vs. on location	122
		4.2.5	Interaction and control.	125
		4.2.0	Monitoring and validation	129

xii Contents

 4.3 Procedures 4.3.1 Equipment 4.3.2 Management 4.3.3 Wizard of Oz 4.3.4 Legal aspects 5 SL corpus representation 5.1 Introduction 5.1.1 The transcription of read speech versus the tion of spontaneous speech 5.1.2 Levels and types of transcription 5.2 Concerning the segmentation and labelling of speec 5.2.1 Background 5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources 			4.2.7	Multi-channel recording)
 4.3.1 Equipment 4.3.2 Management 4.3.3 Wizard of Oz 4.3.4 Legal aspects 5 SL corpus representation 5.1 Introduction 5.1.1 The transcription of read speech versus the tion of spontaneous speech 5.1.2 Levels and types of transcription 5.2 Concerning the segmentation and labelling of speed 5.2.1 Background 5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources 		4.3	Proced	lures)
 4.3.2 Management 4.3.3 Wizard of Oz 4.3.4 Legal aspects 5 SL corpus representation 5.1 Introduction 5.1.1 The transcription of read speech versus the t tion of spontaneous speech 5.1.2 Levels and types of transcription 5.2 Concerning the segmentation and labelling of speed 5.2.1 Background 5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and p representations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources 			4.3.1	Equipment)
 4.3.3 Wizard of Oz 4.3.4 Legal aspects 5 SL corpus representation 5.1 Introduction 5.1.1 The transcription of read speech versus the t tion of spontaneous speech 5.1.2 Levels and types of transcription 5.2 Concerning the segmentation and labelling of speed 5.2.1 Background 5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and p representations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources 			4.3.2	Management	;
 4.3.4 Legal aspects			4.3.3	Wizard of Oz	;
 5 SL corpus representation 5.1 Introduction			4.3.4	Legal aspects	;
 5 SL corpus representation 5.1 Introduction				0 1	
 5.1 Introduction	5	SL c	orpus r	epresentation 146	;
5.1.1 The transcription of read speech versus the trion of spontaneous speech 5.1.2 Levels and types of transcription 5.2 Concerning the segmentation and labelling of speech 5.2.1 Background 5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6.		5.1	Introd	uction	j
tion of spontaneous speech 5.1.2 Levels and types of transcription 5.2 Concerning the segmentation and labelling of speech 5.2.1 Background 5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and p representations			5.1.1	The transcription of read speech versus the transcrip-	
5.1.2 Levels and types of transcription 5.2 Concerning the segmentation and labelling of speed 5.2.1 Background 5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentatio 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and p representations				tion of spontaneous speech	7
5.2 Concerning the segmentation and labelling of speed 5.2.1 Background			5.1.2	Levels and types of transcription)
5.2.1 Background 5.2.2 5.2.2 A caveat 5.2.3 5.2.3 Use of segmented speech databases 5.2.4 5.2.4 The CRIL conventions 5.2.5 5.2.5 Manual segmentation 5.2.6 5.2.6 Automatic and semi-automatic segmentation 5.2.7 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexical information as properties of		5.2	Concer	r_{r} in the segmentation and labelling of speech signals . 150)
5.2.2 A caveat 5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentatio 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.2.1	Background)
5.2.3 Use of segmented speech databases 5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentatio 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.2.2	A caveat	
5.2.4 The CRIL conventions 5.2.5 Manual segmentation 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.2.3	Use of segmented speech databases	
 5.2.5 Manual segmentation			5.2.4	The CRIL conventions	
 5.2.6 Automatic and semi-automatic segmentation 5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script			5.2.5	Manual segmentation)
5.2.7 Segmentation and labelling in the VERBMO 5.3 The levels and types of symbolic representations of 5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.2.6	Automatic and semi-automatic segmentation 153	5
 5.3 The levels and types of symbolic representations of 5.3.1 Recording script			5.2.7	Segmentation and labelling in the VERBMOBIL project 154	ŀ
5.3.1 Recording script 5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources		5.3	The le	vels and types of symbolic representations of speech 155	,
5.3.2 Orthographic transcription 5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.1	Recording script	,
5.3.3 Morphological, syntactic, semantic and prepresentations 5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.2	Orthographic transcription	,
representations			5.3.3	Morphological, syntactic, semantic and pragmatic	
5.3.4 Citation-phonemic representation 5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources				representations	;
5.3.5 Broad phonetic or phonotypic transcription 5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.4	Citation-phonemic representation	;
5.3.6 Narrow phonetic transcription 5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.5	Broad phonetic or phonotypic transcription 159)
5.3.7 Acoustic-phonetic transcription 5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.6	Narrow phonetic transcription)
5.3.8 Physical transcription 5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.7	Acoustic-phonetic transcription)
5.3.9 Prosodic transcription 5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.8	Physical transcription	
5.3.10 Non-linguistic and other phenomena 5.4 List of recommendations 5.4 List of recommendations Part II: Spoken language characterisation 6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources			5.3.9	Prosodic transcription	
 5.4 List of recommendations			5.3.10	Non-linguistic and other phenomena	;
 Part II: Spoken language characterisation 6 Spoken language lexica 6.1 Introduction		5.4	List of	recommendations)
 Part II: Spoken language characterisation 6 Spoken language lexica 6.1 Introduction					
 Part II: Spoken language characterisation 6 Spoken language lexica 6.1 Introduction					
6 Spoken language lexica 6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources	Pa	rt II	: Sp	oken language characterisation 175)
 6 Spoken language lexica 6.1 Introduction		a -	-		
6.1 Introduction 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words 6.1.3 Recommendations on resources	6	Spol	ken lang	guage lexica 176	;
 6.1.1 Lexica for spoken language systems 6.1.2 Lexical information as properties of words . 6.1.3 Recommendations on resources 		6.1	Introd	uction	;
6.1.2 Lexical information as properties of words . 6.1.3 Recommendations on resources			6.1.1	Lexica for spoken language systems	;
6.1.3 Recommendations on resources			6.1.2	Lexical information as properties of words	,
			6.1.3	Recommendations on resources	,

6.2What is a spoken language lexicon?1846.2.1Basic features of a spoken language lexicon184

	6.3	$\begin{array}{c} 6.2.2 \\ 6.2.3 \\ 6.2.4 \\ 6.2.5 \\ 6.2.6 \\ \text{Types} \\ 6.3.1 \\ 6.3.2 \\ 6.3.3 \\ 6.3.4 \end{array}$	Lexical databases and system lexica for spoken language Spoken language and written language lexica Basic lexicographic coverage criteria	e185 187 189 190 191 192 192 193 195
			guage	200
		6.3.5	Recommendations on types of lexical information	202
	6.4	Lexica	$l \ surface \ information $	203
		6.4.1	Orthographic information	203
		6.4.2	Pronunciation information	204
		6.4.3	Prosodic information	209
		6.4.4	Recommendations on lexical surface information	211
	6.5	Morph	ological information	212
		6.5.1	Types of morphological information	212
		6.5.2	Applications of morphology	213
		6.5.3	Recommendations on morphology	216
	6.6	Gram	natical information	216
		6.6.1	Statistical language models	216
		6.6.2	Sentence syntax information	217
		6.6.3	Recommendations on grammatical information	218
	6.7	Lexica	$l \text{ content information } \dots $	219
		6.7.1	Lexical semantic information	219
		6.7.2	Pragmatic information	219
		6.7.3	Idiomatic information	220
		6.7.4	Recommendations on semantic information	221
	6.8	Lexico	$n \ structure$	221
		6.8.1	Spoken language lexicon formalisms	221
		6.8.2	Lexicon architecture and lexical database structure	223
		6.8.3	Lexicon architecture and the structure of lexical	
			databases	225
		6.8.4	Recommendations on lexicon structure	231
	6.9	Lexica	l knowledge acquisition for spoken language \ldots	231
		6.9.1	Stages in lexical knowledge acquisition	231
		6.9.2	Types of knowledge source	232
		6.9.3	Recommendations on lexicon construction	234
	6.10	Outloc	$\mathbf{b}\mathbf{k}$	235
7	Lang	guage m	nodels	237

Contents

7.1	Introd	uction	. 237
	7.1.1	Goals of the chapter	. 237
	7.1.2	Advice to the reader	. 238
7.2	System	n architecture for speech recognition	. 238
	7.2.1	Bayes decision rule	. 238
	7.2.2	Stochastic language modelling	. 240
	7.2.3	Types of language models	. 243
	7.2.4	Why does the trigram language model work?	. 244
7.3	Perple	exity definition	. 245
	7.3.1	Formal definition	. 245
	7.3.2	Implications of the perplexity definition	. 246
	7.3.3	Extensions of the perplexity definition	. 247
	7.3.4	Recommendations	. 247
7.4	Langu	age model smoothing: modelling unseen events	. 248
	7.4.1	Problem formulation	. 248
	7.4.2	Linear discounting and backing-off	. 249
	7.4.3	Linear interpolation	. 252
	7.4.4	Absolute discounting and backing-off	. 252
	7.4.5	Conclusion	. 253
7.5	Multil	evel smoothing for trigram models	. 253
	7.5.1	Problem formulation	. 253
	7.5.2	The full trigram model	. 254
	7.5.3	Practical issues	. 255
	7.5.4	Cache	. 256
	7.5.5	Experimental results	. 257
	7.5.6	Recommendations: m -gram language models	. 261
7.6	Refine	ed language models	. 262
	7.6.1	Equivalence histories and word classes	. 262
	7.6.2	Grammar based language models	. 266
	7.6.3	Recommendations: Refined language models	. 270
7.7	Langu	age models and search	. 270
	7.7.1	Sources of recognition errors	. 270
	7.7.2	Search: Finding the single best sentence	. 271
	7.7.3	Search: Word graphs	. 275
	7.7.4	Recommendations: Language models and search	. 281
7.8	Final:	note: the mathematics of smoothing	. 281
	7.8.1	Linear discounting and backing-off	. 281
	7.8.2	Linear interpolation	. 284
	7.8.3	Absolute discounting and backing-off	. 286
\mathbf{Phy}	sical ch	aracterisation and description	288
8.1	Introd	uction	. 288
	8.1.1	The communication chain	. 289

xiv

8

Contents	xv	
----------	----	--

		8.1.2	Specifications for speech corpora production and speech assessment environments	290
		813	Becommendations on recording strategies	291
	8.2	Basic	notations and terminology	291
	0	8.2.1	Periodic and aperiodic motion	291
		8.2.2	The dB scale	292
		8.2.3	Phones and somes	293
		8.2.4	Analog and digital speech signal representation	293
	8.3	Talker	/ listener descriptors	295
	0.0	8.3.1	General (demographic) descriptors	295
		8.3.2	Task specific descriptors	296
	8.4	Transo	lucer characteristics	301
		8.4.1	Microphones and speech recordings	302
		8.4.2	Parallel recordings	317
	8.5	Enviro	onment characteristics	321
		8.5.1	Artificial vs. natural environment	321
		8.5.2	Auditory environment	322
		8.5.3	Visual environment	330
		8.5.4	Miscellaneous environmental factors	331
	8.6	Record	ling chain characteristics	331
		8.6.1	Signals and systems	332
		8.6.2	Linear distortions	333
		8.6.3	Non-linear distortions	335
		8.6.4	Recommendations on recording chain characteristics .	335
	8.7	Repro	ducibility assurance procedures	336
		8.7.1	Description of recording conditions	336
		8.7.2	References and standards	337
		8.7.3	Recommendations on reproducibility	338
	8.8	Analys	sis and conditioning tools for speech signals $$	339
		8.8.1	Signal analysis and measurement	339
		8.8.2	Signal conditioning	340
		8.8.3	Recommendations on analysis and conditioning tools	
			for speech signals	341
Pε	nrt II	I: S	poken language system assessment	343
9	Asse	essment	methodologies and experimental design	344
	9.1	Introd	uction	344

malou	uction
9.1.1	How to read this chapter
9.1.2	Role of statistical analysis and experimentation in
	Language Engineering Standards (LES)

Contents

	9.2	Statist	ical and experimental procedures for analysing data	
		corpor	a	347
		9.2.1	Statistical analysis	347
		9.2.2	Populations, samples and other terminology	347
		9.2.3	Sampling	348
		9.2.4	Biases	348
		9.2.5	Estimating sample means, proportions and variances	349
		9.2.6	Hypothesis testing	354
	9.3	Experi	mental procedures	360
		9.3.1	Experimental selection of material	360
		9.3.2	Segmentation	363
		9.3.3	Classification	365
	9.4	Assess	ing recognisers	366
		9.4.1	Baseline performance	366
		9.4.2	Progress	367
		9.4.3	Functional adequacy and user acceptance	370
		9.4.4	Methodology	372
		9.4.5	Experimental design	373
	9.5	Assess	ing speaker verification and recognition systems	373
		9.5.1	Sampling rare events in speaker verification and	
			recognition systems	374
		9.5.2	Employing expert judgments to augment speaker ver-	
			ification and assessment for forensic aspects of speaker	
			verification and recognition	374
	9.6	Interac	tive dialogue systems	375
		9.6.1	Wizard of Oz (WOZ)	375
		9.6.2	Dialogue metrics	378
			-	
10	Asse	ssment	of recognition systems	381
	10.1	Introd	uction	381
		10.1.1	Classification of recognition systems	381
		10.1.2	Speech quality and conditions	383
		10.1.3	Capability profile versus requirement profile	384
		10.1.4	Assessment purpose versus methodology	385
	10.2	Definit	ions and nomenclature	385
		10.2.1	The performance measure as percentage	385
		10.2.2	Recognition score	386
		10.2.3	Confusions	388
		10.2.4	Vocabulary	388
		10.2.5	Analysis of Variance design	389
	10.3	Descri	ption of methodologies	389
		10.3.1	Representative databases	389
		10.3.2	Reference methods	390

xvi

Contents	xvii
0 0	

		10.3.3	Specific calibrated databases
		10.3.4	Diagnostic methods with a specific vocabulary 391
		10.3.5	Artificial test signals
	10.4	Param	eters
		10.4.1	Pre-production parameters
		10.4.2	Post-production parameters
		10.4.3	Linguistic parameters
		10.4.4	Recogniser specific parameters
		10.4.5	Assessment parameters
	10.5	Experi	mental design of small vocabulary word recognition 396
		10.5.1	Technical set-up
		10.5.2	Training
		10.5.3	Test procedure
		10.5.4	Scoring the results
		10.5.5	Analysis of results
	10.6	Experi	mental design of large vocabulary continuous speech
		recogn	ition
		10.6.1	Training material
		10.6.2	Development test
		10.6.3	Drv run
		10.6.4	Test material selection
		10.6.5	Evaluation protocol
		10.6.6	Scoring method
11	Asse	ssment	of speaker verification systems 408
	11.1	Presen	tation
		11.1.1	Speaker classification tasks
		11.1.2	General definitions
	11.2	A taxo	nomy of speaker recognition systems
		11.2.1	Task typology
		11.2.2	Levels of text dependence
		11.2.3	Interaction mode with the user
		11.2.4	Definitions
		11.2.5	Examples
	11.3	Influen	cing factors
		11.3.1	Speech quality
		11.3.2	Temporal drift
		11.3.3	Speech quantity and variety
		11.3.4	Speaker population size and typology
		11.3.5	Speaker purpose and other human factors 421
		11.3.6	Recommendations 423
		11.3 7	Example 426
	11 4	Scorin	g procedures 428
	****	~~~~	\mathbf{D} processing of the transformed states

		11 / 1	Natation	190
		11.4.1	Classed and identification	420
		11.4.2	V is the set identification $\cdots \cdots \cdots$	431
		11.4.3		450
		11.4.4	Open-set identification	458
		11.4.5	Recommendations	459
	11.5	Compa	arative and indirect assessment	461
		11.5.1	Reference systems	461
		11.5.2	Human calibration	463
		11.5.3	Transformation of speech databases	464
	11.6	Applic	ations, systems and products	464
		11.6.1	Terminology	465
		11.6.2	Typology of applications	466
		11.6.3	Examples of speaker verification systems	469
		11.6.4	Examples of speaker verification products	470
		11.6.5	Alternative techniques	471
		11.6.6	Conclusions	473
	11.7	System	and product assessment	473
		11.7.1	System assessment	474
		11.7.2	Product assessment	475
		11.7.3	Recommendations	476
	11.8	Forens	ic applications	477
		11.8.1	Listener method	477
		11.8.2	Spectrographic method	478
		11.8.3	Semi-automatic method	479
		11.8.4	Recommendations	479
	11.9	Conclu	isions	480
		0000000		
12	Asse	ssment	of synthesis systems	481
	12.1	Introd	uction	481
		12.1.1	What are speech output systems?	481
		12.1.2	Why speech output assessment?	482
		12.1.3	Users of this chapter	483
	12.2	Toward	ds a taxonomy of assessment tasks and techniques	485
		12 2 1	Glass hox vs. black hox	485
		12 2 2	Laboratory vs. field	487
		12.2.2	Linguistic vs. acoustic	488
		12.2.0	Human vs. automated	488
		12.2.4	Indement vs. functional testing	180
		12.2.5	Global vs. analytic assessment	10 <i>5</i> 400
	19 2	Metho	dology	490 400
	14.0	19 2 1	Subjects	401
		10.0.1	Test procedures	-191 404
		10 2 2	Renchmarks	494 407
		⊥⊿.ಎ.ಎ		431

|--|

12.3.4 Reference conditions	497
12.3.5 Comparability across languages	501
12.4 Black box approach	502
12.4.1 Laboratory testing	502
$12.4.2$ Field testing \ldots	508
12.5 Glass box approach	511
12.5.1 Linguistic aspects	511
12.5.2 Acoustic aspects	518
12.6 Further developments in speech output testing	540
$12.6.1$ Introduction \ldots \ldots \ldots \ldots \ldots \ldots \ldots	540
12.6.2 Long-term strategy: Towards predictive tests	541
12.6.3 Linguistic testing: Creating test environments for lin-	
guistic interfaces	544
12.6.4 Acoustic testing: Developments for the near future	546
12.7 Conclusion: summary of test descriptions	550
12.7.1 SAM Standard Segmental Test	551
12.7.2 CLuster IDentification Test (CLID)	552
12.7.3 The Bellcore Test	553
12.7.4 Diagnostic Rhyme Test (DRT)	554
12.7.5 Modified Rhyme Test (MRT)	555
12.7.6 Haskins Syntactic Sentences	556
12.7.7 SAM Semantically Unpredictable Sentences (SUS)	557
12.7.8 Harvard Psychoacoustic Sentences	558
12.7.9 SAM Prosodic Form Test	559
12.7.10 SAM Prosodic Function Test	560
12.7.11 SAM Overall Quality Test	561
12.7.12 ITU-T Overall Quality Test	562
13 Assessment of interactive systems	564
13.1 Introduction	564
13.1.1 About this chapter	564
13.1.2 Reading guide	565
13.2 Interactive dialogue systems	567
13.2.1 Definitions	567
13.3 Specification and design	571
13.3.1 Design by intuition	573
13.3.2 Design by observation	578
13.3.3 Design by simulation	581
13.3.4 Iterative design methodology for spoken language di-	
alogue systems	591
13.3.5 Readings in interactive dialogue system specification	594
13.4 Evaluation	594
13.4.1 Background	595
	555

xx Contents

13.4.2	Characterisation	. 597
13.4.3	Assessment framework	. 601
13.4.4	Recommendations on evaluation methodology	. 613
13.4.5	Readings in interactive dialogue system evaluation	. 614

Bibliographical references

617

Pa	rt IV	V: SI	poken language reference materials	653
Α	Chai	acter co	odes and computer readable alphabets	654
	A.1	Introdu	uction	654
	A.2	ASCII		654
		A.2.1	Problems	655
	A.3	Multi-I	Byte encodings	656
		A.3.1	ISO 10646	656
		A.3.2	The Unicode standard	656
	A.4	Transla	ation of code tables	657
		A.4.1	Specific vs. general converters	657
		A.4.2	MIME	658
	A.5	Compu	ıter readable phonetic alphabets	658
		A.5.1	IPA subsets	659
		A.5.2	Full IPA alphabet encoding	660
	A.6	Referen	nces	662
в	SAN	IPA con	nputer readable phonetic alphabet	684
	B.1	Introdu		684
		B.1.1	Notation issues	684
		B.1.2	Transcription	684
		B.1.3	Coding	685
		B.1.4	Further languages	687
	B.2	SAMP.	A: Present status	688
	B.3	The ph	nonemic notation of individual languages	691
		B.3.1	Bulgarian	692
		B.3.2	Danish	694
		B.3.3	Dutch	697
		B.3.4	English	699
		B.3.5	Estonian	703
		B.3.6	French	705
		B.3.7	German	707
		B.3.8	Greek	710
		B.3.9	Hungarian	712
		B.3.10	Italian	714

Contents	xxi

		B.3.11 Norwegian	7
		B.3.12 Polish)
		B.3.13 Portuguese)
		B.3.14 Romanian	2
		B.3.15 Spanish	3
		B.3.16 Swedish	j
	B.4	Levels of annotation and extension of SAMPA	7
		B.4.1 SAMPA as a phonemic system	7
		B.4.2 Detailed phonetic or acoustic annotation	7
	B.5	SAMPROSA (SAM PROSodic Alphabet): symbol set 730)
	B.6	References	-
С	SAM	file formats 732	,
	C.1	Introduction)
	C.2	Speech file and associated description file formats	2
	Ċ.3	Label file format)
	0.0	C.3.1 Label file header format	3
		C.3.2 Label file body format	ŧ
		C.3.3 Label types	ŧ
		C.3.4 Label File example	ý
	C.4	Files used in a recording session using EUROPEC	7
	-	C.4.1 Corpus File: CORPUS.DBF	7
		C.4.2 Speaker File: SPEAKERS.DBF)
		C.4.3 Prompt file)
	C.5	Protocol description file)
	0.0	C.5.1 Codes used for presentation protocol (Protocol De-	
		scription File)	2
		C.5.2 Example of Protocol Description File	3
	C.6	Recording conditions file	, j
	C.7	PRIVATE.PEC file	j
П	SAM	recording protocols 747	7
Ľ	D_1	Definition of terms 747	7
	D.1	Classification of general strategies for recording and prompting 748	2
	D.2	D = 21 Recording mode 748	, 2
	ЪЗ	Prompting style 748	, 2
	D.0	D 3 1 Timing strategy 740	,)
	D 4	Becording protocol 740	,)
	D.1	D_{41} Microphone 740	, 1
		D 4 2 Other sensors 750	ì
		D.4.2 Speech data capture 750	,)
		D 4 4 Becording environment 751	'
		D 4.5 Becording mode and prompting style 751	•
		D. 1.0 Recording mode and prompting style 101	•

	D.4.6Recording control75D.4.7Recording procedure75D.4.8Integrity checks75D.4.9Backup procedures75D.4.10Retrieval procedures75D.4.11Calibration75D.4.12Inter site consistency and recording procedure verification75D.4.13Collation of recordings75	52 54 55 55 55 56 56
Ε	SAM software tools75E.1DKISALA (Interactive Semi-Automatic Labelling Software)75E.2ELSA (ESPRIT Labelling System Assessment software)76E.3EUROPEC (European Program d'Enregistrement de Corpus)76E.4PTS (Progiciel de Traitement de Signal)76E.5RESAM76E.6SAMITPRO (SAM Iterative Proportional Fitting)76E.7SAM_REC0 Isolated Word Recogniser76E.8SAM_SCOR (Sam Input Assessment Scoring Software)76E.10SAM_SPEX (Speech Parameter Extractor)76E.11SAMTRA (SAM TRanscription Analysis)76E.13PTM (Parametric Test Manager)76	58 58 59 50 51 52 53 53 55 56 57 57 58
F	EUROPEC recording tool 77	70
G	Digital storage media77G.1 Storage media dimensions77G.1.1 Permanence77G.1.2 Order of access77G.1.3 Other storage media characteristics77G.1.4 Compression77G.2 CD-ROM77G.2.1 Physical CD formats77G.2.2 Logical data organisation77G.2.3 CD-R software77G.2.5 Future developments77G.3 Summary77G.4 References77	73 73 73 74 75 75 76 77 77 78 79
Η	Database Management Systems (DBMSs)78H.1Data model	30 30

ents xxiii

	H.1.1 Hierarchical data model	781 781 784 784 785 785 785 786 786 786
Ι	Speech standardsI.1IntroductionI.2Signal theory recallsI.3Speech researchI.4Computer hardware and softwareI.5ConclusionI.6Appendix: Useful anonymous ftp sites	788 788 788 789 793 794 795
J	EUROM-1 database overview	796
К	Polyphone project overview K.1 Introduction K.2 Reusable resources K.3 The Dutch POLYPHONE corpus K.3.1 Recording workstation K.3.2 Speaker selection K.3.3 The speech material K.3.4 Postprocessing K.4 Use of POLYPHONE in application development K.4.1 Train time table information K.4.2 Phone card services K.4.3 Speaker verification	798 798 799 800 800 800 802 802 802 802 802 805 806
L	 European speech resources L.1 Overview L.2 Criteria for assessment of the situation of Spoken Language Resources L.2.1 Types and specificities of corpora L.2.2 Actors in speech resource production L.3 Summary of the current situation on a per language basis L.3.1 DANISH (Denmark) L.3.2 DUTCH (The Netherlands) 	808 809 810 811 812 812 813

				T7' 1	\ \							01.0
		L.3.3 EN	GLISH (United	Kingdom	1)	•••		• •	• •	•	• •	813
		L.3.4 FR	ENCH (France,	Belgium,	Swit	zerla	nd)	• •	• •	•	• •	814
		L.3.5 GE	RMAN (Germa	ny)				• •				815
		L.3.6 GF	EEK (Greece).									816
		L.3.7 IT.	ALIAN (Italy) .									817
		L.3.8 NC	RWEGIAN (No	rway) .								817
		L.3.9 PC	RTUGUESE (P	ortugal)								818
		L.3.10 SP	ANISH (Spain)									819
		L.3.11 SW	EDISH (Sweder	1)								819
	L.4	General co	nclusions									820
	L.5	Production	1 costs									822
Μ	Tran	scription a	nd documentatio	on conven	tions	for §	Spee	chd	at			824
	M.1	Introducti	on									824
		M.1.1 Po	nts of departure									824
	M.2	Markings	required for scor	ing								825
		M.2.1 Ca	se									825
		M.2.2 Sp	elling									825
		M.2.3 Nu	mber sequences									826
		M.2.4 Let	ter sequences									826
		M.2.5 Pu	nctuation									827
		M 2.6 Mi	spronunciations					• •		•		827
		M 2.7 Un	intelligible word:	 s		•••		• •	• •	•	• •	827
		M 28 Wc	rd fragments			• •		• •	• •		• •	827
		M20 Vo	bal deletions			• •	•••	• •	• •	•	• •	897
		$M_{2.9}$ ve	bai deletions :	e o v orta	• • •	• •		• •	• •	•	• •	021
		M 9 11 Dm	-speech acousti	c events	• • •	• •		• •	• •	•	• •	040
		M.2.11 Pro	sourc annotatio		• • •	• •	••••	• •	• •	•	• •	049
	110		incated waveform	ns	• • •	• •		• •	• •	•	• •	829
	M.3	Assessmen	t (optional)		• • •	• •		• •	• •	•	• •	830
	M.4	Procedure			• • •	• •	•••	• •	• •	•	• •	831
	M.5	Speaker in	formation		• • •			• •	• •	•	• •	831
	M.6	Some lang	uage-specific tra	nscriptio	n opt	ions		• •	• •	•	• •	831
		M.6.1 Fre	nch			• •		• •	• •	•	• •	831
		M.6.2 En	glish			• •		• •	• •	•	• •	832
		M.6.3 Ita	ian			• •		• •	• •	•		833
	— ••	·										0.G /
Ν	The	Bavarian A	rchive for Speec	h Signals								834
	N.1	Introducti	on			• •		• •	• •	•	• •	834
	N.2	General in	formation		• • •	• •		• •	• •	•	• •	834
	N.3	Corpora .										834
	N.4	Corpora in	$_{1}$ preparation									835
	N.5	Other acti	vities									836
	N.6	Cooperatio	ons									836

		C	Contents	xxv
N.7 N.8	Future plans More information	· · ·		836 836
Glossar	ry		8	837
List of	abbreviations		8	851
Index			ę	857

List of Figures

$1.1 \\ 1.2$	The main structure of the EAGLES Group	$\frac{2}{3}$
1.3	Multimodal human-computer interface (HCI) including speech/language input/output	5
1.4	Working structure of the EAGLES Spoken Language Working	19
1.5	A model of the relationship between the applications of spo-	14
	ken language systems and the underlying technology	11
2.1	Spoken dialogue system	31
2.2	Capabilities versus requirements	33
2.3	Specifications versus capabilities	33
2.4	Technology specifications versus application requirements	34
2.5	Result presentation with "speaker selection" procedure	43
2.6	False acceptance versus false rejection	52
2.(Error rate related to speech input duration	99
5.1	Levels of annotation	149
7.1	Bayes decision rule for speech recognition	240
7.2	Illustration of the decision problem for a three-word vocabular	y242
7.3	Perplexity as a function of b for absolute discounting with backing-off.	261
7.4	Perplexity as a function of λ for linear discounting with	
	backing-off.	262
7.5	Bigram as a finite state grammar	268
7.6	Trigram as a finite state grammar	268
7.7	Parsing in the conventional form	269
7.8	Parsing in link grammar form	270
7.9	Example of a time alignment path	272
7.10	Bigram language model and search	273
7.11	Word recombination for a lexical tree	274
7.12	Simplified example of a word graph	276
7.13	Word pair approximation	278
7.14	Example of a word graph	280
8.1	Scheme of the communication chain	289
8.2	Typical polar patterns of various types of unidirectional mi-	
	crophone	305
8.3	Closed room pressure zones	307
8.4	Typical comb filter structure	308

List of Figures xxvii

$8.5 \\ 8.6 \\ 8.7 \\ 8.8$	The minimum recording chain
9.1	Summary of decision structure for establishing what statistical test to use for data 358
92	Activity associated with signal distribution 368
9.3	Activity associated with noise distribution
9.4 9.5	Activity level associated with signal and noise distribution
	speaker interruption patterns
10.1	Simplest set-up for interactive testing of a word recognition system
11.1	A typical ROC curve and its equal error rate
12.1	Relationships among dimensions involved in a taxonomy of speech output evaluation methods

List of Tables

2.1	Confusion matrix		42
2.2	Error presentation		49
2.3	False acceptance versus false rejection		51
2.4	Different types of error		53
2.5	Erlang B carried-traffic table		76
2.6	Erlang B carried-traffic table (continued)		77
2.7	Erlang B carried-traffic table (continued) $\ldots \ldots \ldots$	•	78
4.1	Possible layout for recording session information $\ldots \ldots$		133
5.1	Table of number-to-symbol correspondence for the $I\!\!\!\!\!\!^{\mathbb A} T_{E} X$ font		
	wsuipa11	•	173
6.1	Examples of morphological coverage		190
6.2	Computer readable ASCII orthography for German		205
6.3	Frequently used symbols		226
6.4	VERBMOBIL diacritics		227
6.5	Extract from the VERBMOBIL pronunciation table	•	228
7.1	Effect of trigram model on recognition (Jelinek 1991)		244
7.2	Number of words and sentences in training and test data		
	(vocabulary: about 20000 words)	•	258
7.3	Number of distinct and of singleton events for bigrams and		
	trigrams	•	258
7.4	Perplexities for different language models	•	259
7.5	Algorithm for word clustering	·	265
7.6	One-pass algorithm for word graph construction ("single		0-0
	best" and "word graph")	•	279
7.7	DP algorithm for search through word graph	•	280
10.1	Classification of speech recognition systems		381
10.2	Conditions of speech	•	383
11.1	Verification Phrase Construction for the TI Operational		
	Voice Verification System (after Doddington)		416
11.2	Example of misclassification & mistrust rate scoring from a		
	confusion matrix in speaker identification		437
11.3	Example of confidence rank scoring in closed-set speaker		
	identification		438
11.4	Genuine attempts		447
11.5	Impostor attempts against registered speakers	•	448

$11.6 \\ 11.7 \\ 11.8 \\ 11.9$	Impostor attempts from impostors	$449 \\ 472 \\ 472 \\ 472 \\ 472 \\ 472 \\$
12.1 12.2	Percentage of correct treatment of textual anomalies by CSTR text preprocessor (after Laver et al. 1988, pp. 12–15) Evaluation results for text preprocessor TextScan (after Van	512
12.3	Holsteijn 1993)	513
12.4 12.5 12.6	Some results for the MRT (Logan et al. 1985)	$514 \\ 521 \\ 522 $
12.7	stein and Till 1992)	523 528
12.8	Percentage of correctly recognised emotions-by-rule in Dutch diphone synthesis (two diphone sets, obtained from different speakers) and in human speech (after Vroomen et al. 1993)	535
13.1	Summary of core comparative evaluation metrics	614
A.1	The KLATTBET phonetic alphabet	660
A.2	Machine Readable Phonetic Alphabet (MRPA)	661
A.3	IPA symbol chart	663
A.4	IPA number chart	664
A.5	IPA table (ordered by symbol)	665
A.6	IPA table (ordered by symbol) (continued)	666
A.7	IPA table (ordered by symbol)(continued)	667
A.8	IPA table (ordered by symbol) (continued)	668
A.9	IPA table (ordered by symbol) (continued)	669
A.10) IPA table (ordered by symbol) (continued)	670
A.11	IPA table (ordered by symbol) (continued)	671
A.12	P IPA table (ordered by symbol) (continued)	672
A.13	B IPA table (ordered by symbol) (continued)	673
A.14	IPA table (ordered by symbol) (continued)	674
A.15	iPA table (ordered by number)	675
A.16	iPA table (ordered by number) (continued)	676
A.17	(IPA table (ordered by number) (continued)	677
A.18	3 IPA table (ordered by number) (continued)	678
A.19	PIPA table (ordered by number) (continued)	679
A.20	IPA table (ordered by number) (continued)	680

xxx List of Tables

A.21	IPA table (ordered by number) (continued) $\hdots\dots\dots\dots\dots\dots$ 681
A.22	IPA table (ordered by number) (continued)
A.23	IPA table (ordered by number) (continued)
B 1	Vowels 688
B.2	Consonants
B.3	Two character symbols
B.4	Boundary and prosodic features
B.5	Currently under discussion
B.6	Currently used in French work
G.1	Storage media classification
I.1	Sampling frequencies
M.1	CSELT mapping between letters and graphemes

1.1 Background

This handbook has been produced as a result of an initiative by the Commission of the European Union in February 1993 to launch, under the auspices of the DG XIII Linguistic Research and Engineering (LRE) Programme, a project aimed at accelerating the provision of common functional specifications for the development of large-scale speech and language resources within Europe. The initiative became known as "EAGLES" – the Expert Advisory Group on Language Engineering Standards – and it attracted a budget from the European Commission of approximately 1.25 MECU over a period covering almost three years.

1.1.1 EAGLES objectives

The overall strategic aims of the EAGLES initiative were as follows:

- to produce publicly defined and commonly agreed specifications and guidelines for specific areas of language engineering,
- to bring together industry and academia in an attempt to reconcile the often heterogeneous interests and approaches pursued by the two groups,
- to create focal points of expertise in Europe,
- to complement related European R&D projects falling under the LRE, ESPRIT and EUREKA programmes,
- to enhance the usability, portability and comparability of EU projects' results and thus maximise return on investment in the development of language products and services,
- to contribute to consensus building on an international scale by interacting with national and international standardisation initiatives, and
- to safeguard the multilingual dimension of Europe.

In addition, more specific objectives were defined as follows:

- to promote and accelerate cooperation and consensus building in specific areas of language engineering in Europe,
- to produce prenormative specifications and guidelines for the description and representation of linguistic knowledge and data, including methods for the assessment and evaluation of systems and components which make use of such information,
- to play an active role in the definition, demonstration, evaluation, validation, promotion and dissemination of said specifications, thus furthering harmonisation of the methods and formats used for the encoding and interchange of linguistic knowledge and data,
- to exploit and complement the results achieved in EAGLES members' own R&D activities,

- 2 User's guide
 - to actively seek cohesion with cooperative R&D efforts in Europe and worldwide,
 - to produce a set of guidelines based on a broad consensus among participating organisations, which will be made publicly available and to which European and national R&D activities will be invited to adhere, and
 - to provide input to national and European standardisation bodies active in relevant fields.

1.1.2 EAGLES organisational structure

The activities of the EAGLES initiative were distributed over five specialist technical working groups (see Figure 1.1). These working groups were set up in response to the most urgently felt requirements in advanced language engineering: common methodologies for the creation and interchange of electronic language resources such as text and speech corpora, computational lexicons and grammar formalisms, and the evaluation and quality assessment of language processing systems and components. There was also felt to be a need to attempt to reconcile the needs and practices of the speech and language R&D communities.



Figure 1.1: The main structure of the EAGLES Group

The EAGLES Management Board was constituted from member organisations representing European projects in natural language and speech (such as MULTILEX, PLUS, ACQUILEX, NERC, GENELEX, SAM-A, SUNDIAL, EU-ROLANG, TWB, ONOMASTICA and DELIS) and European associations and coordinating bodies such as ELSNET, ESCA, FOLLI and the European Chapter of the ACL. The Board was chaired by Prof. Rohrer from the University of Stuttgart.

The project funding arrangements were such that, in the early stages, only travel and administrative expenses were covered by the EAGLES initiative. This meant that, whilst some cover was provided by other EU funded projects, the bulk of the costs incurred in the production of this handbook (and the other outputs from the EAGLES initiative) have been met by purely voluntary donations of time and effort on the part of the many contributors involved.

1.1.3 EAGLES workplan

Work towards the EAGLES objectives was conducted in accordance with two phases of activity each of roughly fifteen months duration (see Figure 1.2). Interim recommendations were released to the speech and language communities at the halfway point during 1994.



Figure 1.2: The EAGLES workplan

The initial activity was to conduct a survey of existing resources and standards. This was achieved by contacting people working in the field, identifying priority requirements, assembling working papers, reports and material, identifying relationships with other activities, scoping the EAGLES activities, producing an initial survey report and disseminating it within each Group and externally. From this survey, a set of initial recommendations was disseminated to the communities for comment.

The second phase of the project involved obtaining and evaluating the feedback to the initial recommendations, identifying areas of further development, producing and disseminating a set of final recommendations and obtaining feedback on the final results.

1.2 Spoken Language systems, standards and resources

1.2.1 Spoken Language systems

There is a wide range of technologies which fall under the general banner of "spoken language processing" (SLP) including:

- "automatic speech recognition" ASR (also known as "direct voice input" DVI, and "speech input" SI),
- "automatic speech generation" ASG (also referred to as "direct voice output" DVO, "speech synthesis" SS, and "text-to-speech" TTS),
- "speech input/output" SIO (which includes "speech understanding systems" SUS,
- "spoken dialogue systems" SDS, and "speech-to-speech translation systems" STS),
- "speech coding" (covering wide-band coding at over 4k bps, narrow-band secure voice between 1200 bps and 4k bps, and very-low data-rate speech communications at under 1200 bps),
- "speech analysis or paralinguistic processing" (which includes speaker identification/verification, language identification/verification and topic spotting),
- general speech processing applications such as "speech enhancement" and "voice conversion", and "speech systems technology" (which is concerned with the technology of database recording, corpus transcription, annotation, storage and distribution).

Many of these technologies rely heavily on the availability of substantial quantities of recorded speech material: first, as a source of data from which to derive the parameters of their constituent models (manually or automatically), and second, in order to assess their behaviour under controlled (repeatable) test conditions.

Of course very few spoken language processing applications involve standalone spoken language technology. Spoken language provides an essential component of the more general human-computer interface alongside other input/output modalities such as handwriting, typing, pointing, imaging and graphics (see Figure 1.3). This means that the actions and behaviours of the speech-specific components of a spoken language system inevitably have to be orchestrated with respect to the other modalities and to the application itself by some form of interactive dialogue process (simultaneously taking into account the wide range of human factors involved).

The complexity of the human-computer interface, and the subtle role of speech and language processing within it, has been (and continues to be)

a prime source of difficulty in deploying spoken language systems in "real" applications. Not only are field conditions very different to laboratory conditions, but there has been a serious lack of agreed protocols for testing such systems and for measuring their overall effectiveness.



Figure 1.3: Multimodal human-computer interface (HCI) including speech/language input/output

1.2.2 Standards and resources for Spoken Language systems

As indicated above, the domain of the spoken language technologies ranges from speech input and output systems to complex understanding and generation systems, including multimodal systems of differing complexity (such as automatic dictation machines) and multilingual systems (with applications in different languages, but also, as in speech-to-speech translation systems, integrating processors for more than one language). The definition of defacto standards and evaluation methodologies for such systems involves the specification and development of highly specific spoken language corpus and lexicon resources together with suitable measurement and evaluation tools. These requirements currently still determine considerable differences between spoken and written language in terms of paradigms and techniques of measurement and evaluation, which range from different practical and legal requirements for corpus construction to differences in experimental paradigms for the quality control of working systems. In these areas, the de facto standards are derived from the consensus within the spoken language community on evaluation methods and the resources required for these.

Of course, spoken language technology is still a relatively young area and thus the so-called standards that are discussed here represent only the first rung of the ladder towards the more formal standards which might emerge at a later date. The use of the term "standards" in the R&D community and in the context of this handbook is more usefully interpreted in terms of guidelines and recommended practices. The emergence of more prescriptive actions such as professional codes of conduct, quality marks and formal standards still lies very much in the future.

Nevertheless, the requirement for agreed standards and guidelines pervades all of the links in the spoken language system R&D chain starting from the research community (for algorithm development and benchmarking), to product developers (for performance optimisation), system integrators (for component selection), manufacturers (for quality assurance), sales staff (for marketing), customers (for product selection) and users (for service selection).

Of course, activity in the area of standards and resources for spoken language systems is not new; for many years, the majority of spoken language R&D groups have appreciated the value of sharing recorded speech material and the importance of establishing appropriate infrastructure in terms of standardised tools, research methodology, data formats, testing procedures etc. Indeed, the national research communities in a number of countries have put into place mechanisms for discussing and exchanging such information either as a result of an initiative on the part of the research community itself (for example, the Speech Technology Assessment Group - STAG was set up in the UK under the auspices of the Institute of Acoustics in 1983 and the IEEE operated a similar working group in the USA over ten years ago) or mediated by a central agency (such as GRECO in France and DARPA in the USA). Also, several national standards organisations have become involved, notably the National Institute for Standards and Technology (NIST – formally the National Bureau of Standards) in the USA, the National Physical Laboratory (NPL) in the UK and AFNOR in France.

1.2.2.1 Spoken Language standards and resources in Europe

The most significant activity on spoken language standards and resources in Europe has without doubt been the ESPRIT Speech Assessment Methods (SAM) project which ran from 1987 to 1993 (Fourcin 1993; Winski and Fourcin 1994). The SAM project arose out of the need to develop a common methodology and standards for the assessment of speech technology systems which could be applied within the framework of the different European languages. The definition of the project took place in the context of several ongoing national and international programmes of research including the UK Alvey programme, GRECO in France, COST in Europe and DARPA in the USA.

1.2.2.2 The ESPRIT SAM project

The SAM project was based on a collaboration between almost thirty laboratories in eight different countries: six countries within the EU and two from EFTA. Work was conducted in three interconnected areas:

- speech recognition assessment,
- speech synthesis assessment,
- enabling technology and research.

Within this structure SAM established a set of common tools which have become widely used in a large number of participating and non-participating speech research laboratories. These tools included a reference workstation, a recommended set of protocols for recording, storing, annotating and distributing speech data, and a standard machine readable phonetic alphabet. The SAM reference standard workstation (SESAM) was designed to provide a gateway between one European speech research laboratory and another. The minimum hardware requirements were an IBM PC-AT (or compatible) computer, an analogue interface board (OROS-AU21 or AU22), 1Mbyte of extended memory and a CD-ROM reader. SESAM hosted all SAM software products including EUROPEC, VERIPEC, PTS and ELSA for speech data collection and annotation, EURPAC and SAM_SCOR for measuring the performance of speech recognition systems, and SOAP for measuring the performance of speech synthesis systems.

The first SAM corpus – EUROM-0 – was distributed on a single CD-ROM and contained five hours of speech material. A second corpus – EUROM-1 – used the same standard format with sixty talkers in each of eight languages, speaking phonetically balanced CVC words, number sequences up to 9999 and situationally linked sentence.

1.2.2.3 Other EU projects

In parallel with (and subsequent to) SAM, a number of other EU funded projects have focused on spoken language standards and resources. For example, SQALE was concerned with the assessment of large-vocabulary automatic speech recognition systems across different EU languages and both SUNDIAL and SUNSTAR were directed towards the assessment of multimodal interactive systems.

Other projects with significant outputs in the domain of assessment and resources include ARS, RELATOR, ONOMASTICA and SPEECHDAT, as well as major national projects and programmes of research such as the German VERBMOBIL project. In particular, one of the single, most important achievements of the SPEECHDAT project has been to initiate the creation of the European Language Resources Association (ELRA).
1.2.2.4 The European Language Resources Association

The European Language Resources Association was established in Luxembourg in February, 1995, with the goal of creating an organisation to promote the creation, verification, and distribution of language resources in Europe. A non-profit organisation, ELRA aims to serve as a central focal point for information related to language resources in Europe. It is intended that it will help users and developers of European language resources, as well as government agencies and other interested parties, exploit language resources for a wide variety of uses. It will also oversee the distribution of language resources. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world (such as the LDC – see below).

ELRA membership is open to any organisation, public or private. Full Membership, with voting rights, is available to organisations established in the EU or European Economic Area. Organisations based elsewhere may participate as subscribers. Purely for organisational purposes, members are classified by their chief interest (spoken, written, or terminological resources). The annual membership fee has been set at a level which would encourage broad participation.

1.2.2.5 Spoken Language standards and resources worldwide

At the international level, the NATO Research Study Group on Speech Processing (NATO/AC342/Panel III/RSG10) has, since the late 1970s, provided an effective mechanism for exchanging information on spoken language standards and resources between Canada, France, Germany, the Netherlands, the UK and the USA (Moore 1986). RSG10 was responsible for the first publicly available multilingual speech corpus, and has subsequently released on CD-ROM a database of noises from a range of selected military and civil environments (NOISE-ROM) and related experimental test data (NOISEX).

Also, at each IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) during the 1980s, Janet Baker of Dragon Systems regularly organised a side-meeting to discuss speech databases and opportunities to share such data between different laboratories.

1.2.2.6 COCOSDA

More recently, the International Committee for Collaboration in Speech Assessment and Databases – COCOSDA – was established in 1990 to encourage and promote international interaction and cooperation in the foundation areas of Spoken Language Processing (Moore 1991). COCOSDA provides a forum for international action and discussion and gives platforms for groups of workers to exchange information and to set up collaborations in the field of Spoken Language Engineering. Very many of the world's leading workers are amongst it members and the group discussions are open and unconstrained by any special interests. Meetings take place annually as a satellite event to one of the major international conferences.

1.2.2.7 The Linguistic Data Consortium

In the US, the Linguistic Data Consortium (LDC) was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that, in 1995, included about 65 companies, universities, and government agencies. An initial grant of \$5 million from ARPA amplified the effect of contributions (both of money and of data) from the broad membership base, so that there is guaranteed to be far more data than any member could afford to produce individually. In addition to distributing previously-created databases, and funding or co-funding the development of new ones, the LDC has helped researchers in several countries to publish and distribute databases that would not otherwise have been released.

The operations of the LDC are closely tied to the evolving needs of the research and development community that it supports. Since research opportunities increasingly depend on access to the consortium's materials, membership fees have been set at affordable levels, and membership is open to research groups around the world. Although US government investment in LDC database development is continuing, a significant fraction of the consortium budget comes from membership fees. These fees are now adequate to support the central staff organisation, pay database publication costs and underwrite some database creation.

1.3 The EAGLES Spoken Language Working Group (WG5)

The Spoken Language Working Group (WG5) was constituted from nine senior members of the European spoken language R&D community. The selected individuals represented a range of industrial, academic and government interests and each had direct expertise in the area of spoken language standards and resources. The nine members of the Working Group were:

- Prof. Roger K. Moore (Chairman) DRA Speech Research Unit, Malvern, UK
- Dr. Richard Winski (Host) Vocalis, Cambridge, UK
- Prof. Dafydd Gibbon (Rapporteur) Fakultät für Linguistik und Literaturwissenschaft, Bielefeld, Germany
 Dr. John McNaught (Coordinator's Representative)
- Centre for Computational Linguistics, UMIST, Manchester, UK

- Dr. Giuseppe Castagneri CSELT, Turin, Italy
- Prof. Jean-Marc Dolmazon INPG-ICP, Grenoble, France
- Dr. Norman Fraser
- Vocalis, Cambridge, UK
- Prof. Louis Pols
 Institute of Phonetic Sciences, University of Amsterdam, Netherlands
- Prof. Hans Tillman Institut fur Phonetik und Sprachliche Kommunikation, Munich, Germany

From the outset of the EAGLES initiative, it was clear to the Spoken Language Working Group that very substantial resources already existed in regard to established spoken language corpora and widely accepted systems of data description. Also available were integrated systems of assessment and, in some areas, detailed methods of evaluation. In particular, the prior work of the ESPRIT Speech Assessment Methods (SAM) project had already laid down very substantial groundwork in these areas.

There was, however, an urgent need to provide a central focus for the consolidation and appropriate promulgation of these developments – a "handbook" – and the EAGLES initiative provided a unique framework for achieving this in a European context.

As a consequence, the overall objectives of the EAGLES Spoken Language Working Group (WG5) were set down as follows:

- to consult widely with the spoken language science, research, technology and application community,
- to provide a focus for liaison with other national and international bodies in the field,
- to evaluate existing resources and methodologies,
- to identify areas of consensus in respect of spoken language resources and standards,
- to facilitate interchange and cooperation between the speech and natural language communities, and
- to communicate the results in the form of a "handbook of standards and resources for spoken language systems".

The main technical topics addressed by the Spoken Language Working Group were spoken language resources, systems and terminology (Moore 1994b; Winski et al. 1995).

In the resources area, the Group considered that what was required was a review of contemporary and planned national and international corpora, a catalogue of existing speech data archives and a listing of existing distribution centres. The Group also felt that common protocols, formats and tools should be collated covering all of the important aspects of the design, specification, collection, representation, storage and distribution of spoken language corpora.

In the spoken language systems area, the Group identified two important subtopics: system specification (including requirements definitions, interface standards, dialogue engineering and multimedia) and system assessment (covering methodologies for spoken language input assessment, spoken language output assessment, interactive systems and other speech technologies such as speaker verification).

For spoken language terminology, the Group defined the following activities as being required: the creation of an initial word list, it's extension and/or reduction by members of the Group, the definition of each headword and the addition of glosses in major European languages. However, the Group noted that the funding resources available within the EAGLES initiative were unlikely to be sufficient to complete this important task.

1.3.1 Subgroups of the EAGLES Spoken Language Working Group

Organisationally, the work of the Spoken Language Working Group was performed by a number of Subgroups each focussing on specific technical aspects of the area (see Figure 1.4) and each drawing in expertise from outside the main Working Group.

The structure of the Subgroups was designed to parallel the contents of the planned handbook as it was felt that this would not only provide an effective working structure, but would also simplify the management of the overall activity. In general, each Subgroup had three members, one of whom was a member of the main Spoken Language Working Group.

1.3.2 Relationships with the other EAGLES Working Groups

Within the overall EAGLES initiative, there were obvious technical overlaps between the activities of the five Working Groups and, more specifically, between the Spoken Language Working Group and all four other Working Groups (see Figure 1.1). The exchange of information between groups was encouraged, therefore in the first instance, selected members of the Spoken Language Working Group also participated in other Groups. Also, in view of the very specific commonalities that existed between the Spoken Language Working Group and the Lexicon and Corpus Working Groups, common "Cross-Groups" were established.

1.3.3 Workshops

The objectives of the Spoken Language Working Group were progressed through a series of meetings and workshops which took place during the course of the project. Each event involved members of the main Group and selected Subgroups. The primary workshops took place in London (24th– 25th February 1993 and 5th July 1993), Cambridge (1st–2nd November



Figure 1.4: Working structure of the EAGLES Spoken Language Working Group

1993), Paris (19th–20th January 1995), London (9th–10th March 1995), Lisbon (24th–25th May 1995) and London (4th–5th December 1995).

The initial workshops focused on the development of a detailed structural outline of the proposed handbook and activities were initiated in all of the identified technical areas in parallel. Subsequent workshops reflected the adoption of a more pipelined approach and this enabled the more well developed technical areas to be progressed to completion at an earlier stage. The penultimate workshop in Lisbon was concerned with finalising the overall style and content of the handbook and involved over thirty participants (two-thirds from university research laboratories and one-third from industrial R&D establishments).

1.3.4 Production of the handbook

At the start of the EAGLES initiative, it was intended that all authoring activities would involve voluntary effort on the part of a large number of individual contributors. However within the Spoken Language Working Group, it had become apparent by the halfway stage in the project that this strategy would not guarantee that the agreed comprehensive plan for the handbook would in fact be realised. Therefore, in July 1994, a change in the funding arrangements was negotiated such that identified individuals could be paid a modest fee for acting as technical authors for the main chapters of the handbook.

These revised arrangements were a great success in unblocking the authoring log-jam and an initial draft of the handbook was completed in October 1994. This "interim" handbook covered 50% of the planned topics and was circulated widely throughout the international spoken language R&D community (see below).

The remaining topics were addressed by the Working Group during the second half of the project and a first draft of the full handbook was completed in September 1995. This was presented to the community for the first time at EUROSPEECH'95 (and at the following COCOSDA meeting).

1.3.5 Consultation with the R&D Community

From the beginning, the Spoken Language Working Group placed great emphasis on the need to reflect the wide (and possibly disparate) views of the spoken language R&D community at large. Therefore, the activities of the Group were made public at every opportunity (for example, at the annual meetings organised by the International Coordinating Committee on Speech Databases and Assessment – COCOSDA). Also, specific consultation periods were established between November 1993 and July 1994, and from October 1994 to May 1995 in which draft documentation was made available publicly on the Internet.

An additional consultation was coordinated at the international level by the central EAGLES administration. Over forty institutions worldwide were sent copies of the interim handbook for formal review. These included European industries such as Philips, Hewlett-Packard, Daimler-Benz, Siemens, GEC, Vecsys, Alcatel, Dragon UK, MATRA and ENSIGMA; European PTTs such as Telefonica, British Telecom, Telia, Jutland Telephone, CNET and the Dutch PTT; European research institutions such as ENST, University of Valencia, IDIAP, KTH, University of Essex, University of Patras, University of Catalunya, University of Leeds and Cambridge University; non-European industries such as Dragon Systems, ETL, IBM, VPC, Entropic and Apple; non-European research institutions such as OGI, University of Sydney, Uni-

versity of Tsukuba, Australian National University, University of Berkeley and MIT; and important spoken language resource centres such as the Linguistic Data Consortium (LDC) and the US National Institute of Science and Technology (NIST).

The feedback received from this formal review process was incorporated into the activities of the Group and thence into the handbook itself.

1.4 Overview of the handbook

Previous work on standards and evaluation within the spoken language community have lead to an initial documentation of existing practice which is relatively comprehensive but in many respects heterogeneous and widely dispersed. The purpose of this handbook, therefore, is to collect and catalogue this material within a single document. That is not to say that the handbook is recommending or defining a single European standard, rather, it points to contemporary working practices and *de facto* standards where they already exist.

The handbook has been realised as a series of necessarily interrelated chapters, where each chapter provides some introductory background (including definitions of basic terminology) and concise summaries of common approaches, including alternatives, where these exist. Factors pertaining to recommended approaches are outlined, and preferred methods are identified wherever possible.

The overall style of the handbook is to focus, wherever possible, on clear straightforward "recommendations" supported by appropriate overviews, justifications, exemplifications and reference material. Also, each chapter is intended to be somewhat independent of the others, so that the handbook can appear in its final published form not as a single library volume, but as a set of practical paperbacks and, for convenience in reference to specific points, a fully linked hypertext version.

Clearly, in an exercise of this magnitude, harmonisation all of the key concepts cannot be guaranteed. Not only will the reader come across occasional terminological inconsistencies but it is also possible that some recommendations may be in direct conflict with each other. Such circumstances are probably not errors, but a direct consequence of the lack of concordance that has been attempted so far in the spoken language technology R&D community. Subsequent revisions of the handbook will attempt to resolve these issues.

1.4.1 Intended readership

It is intended that the handbook should provide an essential reference work useful to a wide range of laboratories which are concerned with almost any aspect of spoken language technology. In particular, in addressing the production of the handbook, the Working Group has kept in mind that the potential readership should include:

- research workers and system developers who require convenient access to an organised body of specific reference material,
- workers in other countries who require access to well-documented common practice in central Europe,
- newcomers to the field who require introductory material, primarily research workers in related disciplines and students, and
- corporate end-users of spoken language technology, who need to specify, procure or integrate system components, and who require guidance related to system specification and assessment.

The handbook is not intended to be a textbook about state-of-the-art algorithms and techniques in spoken language technology. However, there are dangers involved in the simplistic use of a handbook without a good understanding of the methods and principles involved. A reader who wishes to find out more along these lines can find appropriate tutorial material in a number of relevant books (Ainsworth 1988; Allerhand 1987; Bloothooft et al. 1995; Bristow 1984, 1986; Holmes 1988; House 1988; Lea 1980; Linggard 1985; Mariani 1989; Roe and Wilpon 1994; Rowden 1992; Witten 1982), the proceedings of the major conferences on spoken language processing such as the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), the European Conference on Speech Communication and Technology (EUROSPEECH), the International Conference on Spoken Language Processing (ICSLP), the International Congress of Phonetic Sciences (ICPhS), and journals such as Speech Communication (North-Holland), Computer Speech and Language (Academic Press), the Journal of the Acoustical Society of America, and Natural Language Engineering (Cambridge University Press).

Likewise, although it is intended that the handbook should provide support to non-experts, it should be noted that there is a limit to what can be sensibly achieved on a "DIY" (do-it-yourself) basis. For example, some of the areas touched on in the handbook (room acoustics, for example) are huge fields in themselves. Hence, if detailed guidance is required, it is important to realise that there are a number of individuals and companies who already provide technical consultancy and services in such areas.

1.4.2 Scope

The scope of the handbook fundamentally addresses the resources required for specifying, developing and evaluating spoken language technology components, including automatic speech recognition, speaker recognition and speech synthesis, which themselves are integrated to form interactive systems such as spoken dialogue systems. There is an emphasis upon the

design, collection, representation, characterisation, storage and distribution of speech corpora, as well as upon assessment methodologies for the component technologies and integrated systems.

The handbook is essentially divided into four main parts. The first part (Chapters 2 to 5) is concerned with the design of spoken language systems and addresses spoken language resources (the design, collection, characterisation and annotation of corpora). The second part (Chapters 6 to 8) is concerned with spoken language characterisation (spoken language lexicon design, language models, and physical characterisation). The third part (Chapters 9 to 13) covers assessment methods (for recognition, synthesis, verification and interactive systems). The fourth part is a substantial body of reference material.

1.4.2.1 Spoken Language system design and specification

One of the difficulties which arises from the complexity of the humancomputer interface (HCI) and the position of spoken language within it, is that people concerned with implementing applications are unable to select appropriate HCI components (such as automatic speech recognisers, for example). This arises not just from a lack of standardised evaluation criteria for system components but also from a lack of clear understanding of the implications on overall performance of the performance of each system component.

One possible model for understanding the relationship between spoken language system applications and the corresponding technology is illustrated in Figure 1.5. The key notion which sets it apart from previous models developed by the spoken language R&D community is that it not only focuses on the fact that there are many factors which influence the performance of spoken language systems and that it is necessary to distinguish between "capabilities" and "requirements", but it also emphasises that the purpose of introducing spoken language technology into an application is to achieve the appropriate operational benefits. It is only when all of these features become properly integrated into agreed methods for spoken language system assessment that it will be possible to arrive at a meaningful (and comprehensive) definition of the "suitability" of particular technologies for particular applications.

The model shown in Figure 1.5 indicates clearly that successful implementation of spoken language systems depends only indirectly on the technical features of the system components and on the operational benefits being sought in the applications themselves. What is more important is to develop a process for converting technical features into technical and operational capabilities, and for converting operational benefits into operational and technical requirements. These processes were felt by the Working Group to be so important to the system design process (and hence to the success of



Figure 1.5: A model of the relationship between the applications of spoken language systems and the underlying technology

the technology in the market place), that a chapter outlining design issues should take pole position at the front of the handbook (Chapter 2).

1.4.2.2 Spoken Language resources

Broadly speaking, the spoken language R&D community can be partitioned into two main interest groups: those concerned with "speech science" and those concerned with "spoken language technology". In the main, speech science is the domain of phonetics, linguistics and psychology, and spoken language technology is the domain of engineering, computer science, mathematics and AI. Both areas have a strong need for significant quantities of both transcribed speech data (orthographic, phonetic, prosodic etc.) and digitised acoustic speech recordings (together with the means for accessing selected subsets of the material using the relevant transcriptions and annotation).

Three types of recorded speech are typically of interest (Moore 1992b):

- analytic-diagnostic material which is of primary importance to progress in basic speech science and which is specifically designed to illuminate specific phonetic and linguistic behaviour (for example, lists of all consonantvowel-consonant syllables in a given language);
- 2. general purpose material which includes vocabularies which are either common or which are typical of a wide range of speech technology applications (for example, alpha-numeric words or standard control terms);
- 3. task-specific material which reflects different levels of formalised spoken monologue/dialogue within constrained discourse domains (for example, train timetable enquiries).

Clearly general purpose speech corpora are easy to collect and are useful in a general sense but, of course, they have only limited practical value. On the other hand, although task-specific corpora can be time-consuming to collect and are only relevant to a specific domain, they are obviously directly useful for the purposes of commercial exploitation. Diagnostic corpora are time consuming to design, but they are extremely useful for research purposes and, in the long term, could prove to be the most valuable resource for spoken language R&D.

At the current time there is a growing requirement for recorded speech which is in some sense more "natural" than the so-called "lab-speech" that has been normally collected and studied up to now. This is true for all three types of material identified above. In this context a range of different speaking styles are now of interest: read speech – including talkers with different amounts of formal training and familiarity with the subject matter, spontaneous speech arising from a directed monologue, spontaneous speech arising from a dialogue between human interlocutors, spontaneous speech arising from simulated human–computer interaction – using the socalled "Wizard of Oz" protocol, and spontaneous speech arising from "real" human–computer interaction.

These issues (and the technology required for acquiring spoken language data) are presented in depth in the handbook chapters on spoken language corpus design (Chapter 3) and collection (Chapter 4).

Of course, recorded spoken language data is, in itself, of limited value; the raw acoustic signal needs to be associated with the appropriate phonetic and linguistic transcripts. This is achieved by "annotating" the data with markers which make such relationships explicit and which provides the means by which the data can be accessed, thereby facilitating the organised study of the data and both automatic parameter estimation and assessment for spoken language systems. These issues are dealt with in the chapter on spoken language corpus representation (Chapter 5).

An important linguistic component of any spoken language corpus, and a key feature of a spoken language system, is the set of words that are employed and their associated properties (such as information about pronunciation, grammatical and semantic features) – the "lexicon". This area is treated in the chapter on spoken language lexica (Chapter 6).

Another key linguistic aspect of spoken language which has particular relevance in spoken language technology systems, is concerned with "language modelling" (Chapter 7).

As well as the linguistic characterisation of spoken language corpora described above, there is also a need to be able to characterise such data from an acoustical and electrical perspective. All aspects of the recording chain become important, from the nature of the recording environment, through the types of microphones or headphones that might be used, to issues such as methods for calculating the signal-to-noise ratio. These factors are presented in the chapter on the physical characterisation of spoken language corpora (Chapter 8).

1.4.2.3 Assessment of Spoken Language systems

In the assessment of spoken language systems it is possible to distinguish three main methodologies: live "field" trials, laboratory-based tests and system modelling paradigms (Moore 1992a). The first of these of course is likely to provide the most representative results but, from a scientific point of view, there are likely to be a number of uncontrolled conditions and this limits the degree of generalisation that can be made from application to application. Field trials also tend to be rather costly operations to mount. Laboratory testing is per force more controlled and can be relatively inexpensive, but the main problem is that such tests may be unrepresentative of some (possibly unknown) key field conditions and give rise to the observed large difference between performance in the laboratory and performance in the field. The third possibility, which is itself still the subject of research, is to model the system (and its components) parametrically. In principle, this approach could provide for a controlled, representative and inexpensive methodology for assessment but, as yet, this area is not sufficiently well developed to be useful.

Also, the term "assessment" covers a range of different activities. For example, a suitable taxonomy of assessment activities should include:

- "calibration" (does the system perform as it should),
- "diagnosis" (how well does the system perform under parametrically controlled conditions),
- "characterisation" (how well does the system perform over a range of diagnostic conditions),
- "prediction" (how well will the system perform under different conditions) and
- "evaluation" (how well does the system perform overall).

Of all these, the last – evaluation – has received a the bulk of the attention in spoken language systems assessment.

Given the complexity of the human-computer interface discussed above, it is clear that assessment protocols are required which address a large number of different types of spoken language system. For example, such systems range from laboratory prototypes to commercial off-the-shelf products, from on-line to off-line systems, from stand-alone to embedded systems, from subsystems to whole systems and from spoken language systems to spoken language based HCI systems.

The majority of research in the area of spoken language system assessment has concentrated on evaluating system components (such as measuring the word recognition accuracy for an automatic speech recogniser, for example) rather than overall (operational) effectiveness measures of complete HCI systems. Since the publication of the NBS guidelines in 1985, there have been considerable developments at the international level. In Europe, the ESPRIT SAM project established a standard test harness for both recognisers and synthesisers and in the US a very efficient assessment paradigm has been funded by the Advanced Projects Research Agency (ARPA) which included an efficient production line of "hub and spoke"-style experiments involving the coordinated design, production and verification of data, distribution through the LDC, and with NIST responsible for the design and administration of tests and the collation and analysis of the results.

These activities point strongly to the importance of establishing appropriate "benchmarks", either through the implementation of standard tests, or by reference to human performance or to reference algorithms.

Throughout these issues, it is vitally important that the relevant practitioners are fully competent in the process of experimental design and in the understanding of key issues such as statistical significance. For these reasons, the handbook specifically includes a chapter on this (Chapter 9) at the front of the chapters on assessment.

The chapter on experimental design is followed by chapters which cover the assessment of the three main component technologies: automatic speech recognition (Chapter 10), speaker verification (Chapter 11) and speech synthesis (Chapter 12). These are followed by a chapter concerned with the assessment of interactive spoken language systems (Chapter 13).

1.4.2.4 The reference material

The handbook is structured such that the supporting material for each chapter has been separated from the main text and collated to form a substantial body of reference material spanning all aspects of spoken language standards and resources. The main reference materials covered are:

- Character codes and computer readable alphabets
- SAMPA computer readable phonetic alphabet
- SAM file formats
- SAM recording protocols

- SAM software tools
- EUROPEC recording tool
- Database management system (DBMS) guide
- Speech standards review
- EUROM-1 database overview
- POLYPHONE project overview
- European speech resources
- Transcription and documentation conventions for Speechdat
- The Bavarian archive for speech signals

1.4.3 The main chapters of the handbook

For general orientation purposes, the reader is recommended to refer to the chapters on system design and specification (Chapter 2) and assessment methodologies and experimental design (Chapter 9).

1.4.3.1 System design

The chapter on "system design" is specifically aimed at potential users of spoken language technology (such as system designers or technology procurers) who need to know how to relate the technical features of the technology to the operational benefits they are seeking to achieve. It is intended that this chapter should be able to help such users to communicate effectively with the technologists and technology suppliers, to give guidance as to what questions they should ask, and to provide a means for specifying their requirements in a way which is meaningful to themselves and to the technologists.

The chapter starts with an introduction to the difference between a system's "capability profile" and the requirements of a given application. This is followed by an enumeration of the many and varied factors which influence the performance of the types of spoken language systems covered by the rest of the handbook. Automatic speech recognition systems are treated first, and over twenty factors are presented which range from aspects such as variability in the fluency of the speaker through to variability in the characteristics of telephone handsets. This is followed by a discussion of the different configurational possibilities for speaker verification/identification systems and a brief description of the key facts of speech synthesis systems. Interactive voice systems are introduced and the importance of error recovery strategies is identified.

The chapter goes on to outline key issues associated with the software and/or hardware aspects of the system platform, and highlights the possibilities for system simulation and prototyping, as well as a variety of practical matters ranging from the physical interface between a spoken language system and the host application to dealing with multilinguality.

1.4.3.2 Spoken Language corpus design

The chapter on "spoken language corpus design" is targeted not only at users of speech corpora within the domain of spoken language technology but also to use in other areas such as sociolinguistics, language learning and pathology. It starts with a discussion of the most important differences between written and spoken language data, and then presents examples of the many application areas which require access to spoken language corpora. The second half of the chapter describes how to specify a spoken language corpus, first in terms of the required linguistic content and, second, in terms of the number and types of speakers involved. The latter issue is dealt with in some detail, and relevant speaker characteristics are covered which, among many other things, include the age and sex of each speaker, their smoking and drinking habits, and whether or not they have received any professional speech training.

1.4.3.3 Spoken Language corpus collection

The chapter on "spoken language corpus collection" concentrates on the practical aspects of collecting spoken language material. In the first part, the dimensions of data collection are described which cover different recording scenarios such as studio versus location recording, or interviews versus read material, for example. It is also pointed out that important data about spoken language may be collected from sensors other than a microphone, for example by means of multi-channel recordings of signals derived from laryngography, electropalatography or NMR (Nuclear Magnetic Resonance) imaging.

The second part of this chapter contains recommendations for the actual collection of spoken language data covering the necessary equipment and the data management protocols needed. The legal aspects of recording arbitrary spoken language material are also discussed and appropriate recommendations given. It is the intention that the recommendations contained within this chapter should enable any reasonably competent person to establish a suitable recording environment that will deliver data in a controlled manner and to an acceptable level of technical quality.

1.4.3.4 Spoken Language corpus representation

The chapter on "spoken language corpus representation" describes how, to be of value, a set of "raw" speech recordings needs to be augmented with symbolic annotation covering a range of phonetic and linguistic levels of description. The transcription of spoken language data is discussed (including problems which arise with spontaneous speech or overlapping speech in dialogues), and mechanisms for segmenting and labelling the data are described. This is followed by an extensive presentation of the many possible representational structures ranging from simple orthography, through detailed low-level acoustic-phonetic analysis, to prosodic transcription and other non-linguistic phenomena (such as hesitations or acoustic non-speech events, for example).

1.4.3.5 Spoken Language lexica

The chapter on "spoken language lexica" provides a framework for relating concepts such as the creation of lexica for specific applications, the transfer of lexical resources from one application to another and the automation of these processes. The chapter covers topics such as the basic features of spoken language lexica, the types of information contained within a spoken language lexicon (such as surface, morphological, grammatical, semantic and pragmatic information), lexicon structure (including appropriate formalisms), lexical access and lexical knowledge acquisition (from dictionaries, for example).

1.4.3.6 Language models

The chapter on "language models" is different from the other chapters in that it is more concerned with details of techniques and algorithms. This is because of the central role language modelling plays in spoken language systems and in characterising a spoken language corpus. The chapter covers the different formalisms involved, the definition of the key concept of "perplexity" and a range of practical schemes for developing high quality language models.

1.4.3.7 Physical characterisation and description

The chapter on "physical characterisation and description" is essentially concerned with the non-linguistic aspects of a spoken language corpus. This includes such features as the characteristics of talkers and listeners, the recording environment, the transducer(s) and any communications channel. It also deals with "reproducibility assurance procedures"; that is, recommendations for ensuring the integrity of the data (for example, calibration techniques and the use of reference signals).

1.4.3.8 Assessment methodologies and experimental design

The chapter on "assessment methodologies and experimental design" is intended to provide general guidance to all practitioners in the field in matters relating to formal methods for designing and executing statistically significant experiments and for the meaningful interpretation of experimental results. This relates both to the design of representative spoken language corpora and to the evaluation of spoken language systems.

1.4.3.9 Assessment of recognition systems

The chapter on the "assessment of recognition systems" presents information on the substantial amount of work that has been done in this area over the past years. The chapter starts with a classification of different recognition systems and then introduces various performance measures. A taxonomy of different assessment methodologies is described ranging from the straightforward use of spoken language corpora, to more diagnostic methods and artificial test signals. This is followed by a discussion of the parameters which affect performance including those which affect the speaker (such as workload stress or noise) and those which affect the recogniser (such as noise).

The second half of the chapter provides recommendations on testing procedures for two main classes of speech recognition system: the smallvocabulary isolated-word recogniser and the large-vocabulary continuous speech recogniser. In both cases, attention is given to the training of the system, the test procedures and scoring and analysing the results.

1.4.3.10 Assessment of speaker verification systems

The chapter on the "assessment of speaker verification systems" opens by presenting a taxonomy of system types in which the difference between identification and verification is made clear, and issues such as text-dependency are illuminated. This is followed by an analysis of the factors which influence the performance of speaker recognition systems and the set of recommended scoring procedures which should be used. The chapter concludes with some specific points concerning the forensic use of speaker recognition systems.

1.4.3.11 Assessment of synthesis systems

The chapter on the "assessment of synthesis systems" starts with a taxonomy of assessment task and techniques, distinguishing, for example, between laboratory and field assessment, and between human judgements and automatic testing. A methodology is then presented covering the choice of subjects for listening experiments, the required test procedures and suitable benchmarks and reference conditions. Recommendations are made for "black box" testing of overall output quality, and for "glass box" testing at many detailed levels of analysis. A taste is also given to future developments in synthesis evaluation.

1.4.3.12 Assessment of interactive systems

The chapter on the "assessment of interactive systems" presents recommendations for the specification, design and assessment of interactive systems in which spoken language dialogue plays a major part. After defining different types of dialogue system, the chapter describes in some detail the "Wizard of Oz" paradigm for system simulation and the central role it plays in the design and assessment of interactive systems. The chapter goes on to address methods for characterising dialogue systems, tasks and users, and presents an assessment framework which includes high-level metrics such as correction rate and transaction success.

1.5 The current state of play

The current handbook cannot be considered a final or complete statement of guidelines and recommendations as agreed by the EU spoken language technology community for the following reasons:

- it may have serious omissions,
- some chapters have not been consulted as fully as others,
- it does not cover the full range of spoken language technologies identified earlier, and
- it may go out of date quite quickly due to the speed of technological progress.

Nevertheless it is expected that the present work substantially reflects the community position on a large range of relevant topics, and will prove to be an important interim working document for the provision of commonly agreed working standards and ultimately, where appropriate, may support progression of these *de facto* conventions and practices towards formal representation.

1.6 Possible future actions

1.6.1 Revision and completion of existing documentation

The presently available documentation on spoken language resources, standards and evaluation methodology contains gaps and required fuller consultation on some of the more recently produced material. Several areas, including corpus collection and lexical database techniques and tools as well as the evaluation methodology for complex systems, require updating and additions in the light of recent developments. More precise user targeting is required, with an explicit distinction in information granularity between management/planning and laboratory/project user levels.

1.6.2 Extended survey of existing practice

Industrial participation has so far been considerable, but the coverage of opinion within the field needs to be extended on a broader basis than has so far been possible. First, a further in-depth survey should be made of the requirements of industrial developers and users. Second, a survey of resources and needs in Eastern Europe and the Newly Independent States

formerly in the Soviet Union is required. Equally important is coverage of results of Fourth Framework Programme projects.

1.6.3 Extension of language base

Existing documentation covers the main languages of the European Union, and definition of standard representation techniques for transcription and signal annotation of other languages is urgently required. Of increasing interest in this respect are the languages of Eastern Europe.

1.6.4 Terminology

Although part of the original plan, it was realised that very little attention could be given to this vitally important area simply due to the lack of available resources to fund the detailed work that would need to be undertaken. Some groundwork has been done, but a significant effort is needed to bring it to completion.

1.6.5 Move to prescriptive recommendations

Most of the recommendations put forward in the current handbook are based on *de facto* standards and simply describe the current working practices in spoken language technology R&D. In any future activity, it would be possible to move towards a more prescriptive framework, in which serious consideration is given to recommending particular methods and techniques over and above some others. Clearly this would require a continued commitment to the process of community consultation and feedback.

1.6.6 Publication and dissemination

The available documentation requires new dissemination and publication concepts in line with recent developments in the use of new media and broadband networks. Efficient development and production techniques for different modes of publication and dissemination of complex documents in conventional and hypertext form are required. Legal aspects of accessibility of resources and documentation need to be addressed.

1.6.7 Coordination with other bodies

The relation between European standardisation and evaluation work and European associations such as ELRA, as well as with national spoken language archives and validation centres, requires further study and negotiation.

Some of the results of core work in spoken language which is of secondary value to written language work, such as pronunciation transcriptions for lexica and dialogue corpora, are available as a service to written language groups. However, in addition to the separate consolidation of work in the

two complementary areas, joint consultation will be required in the foreseeable future on complex systems such as automatic dictation systems or speech to speech translation systems.

It is also the case that the mere existence of the current handbook presents a worthwhile starting point for the negotiation of more formal standards through the mechanisms of national and international standards authorities.

1.7 Contact points

An initiative of this nature can only succeed with the full backing of the R&D community. Already, many individuals have committed themselves to making a contribution to the activities of the Group. Inputs are still welcome from all corners of the field. Potential contributors can either contact the following individuals, or access the central EAGLES team (see below).

- Chairman: Roger K. Moore DRA Speech Research Unit St. Andrews Road Malvern, Worcs, WR14 3PS United Kingdom Tel/Fax: +44 1684 89 4091/5103 email: moore@hermes.mod.uk
- Rapporteur: Dafydd Gibbon Universität Bielefeld Fakultät für Linguistik und Literaturwissenschaft P 10 01 31 33501 Bielefeld Germany Tel/Fax: +49 521 106 3510/2996 email: gibbon@spectrum.uni-bielefeld.de
- Host: Richard Winski Vocalis Ltd. Chaston House Mill Court Gt. Shelford Cambridge United Kingdom Tel/Fax: +44 1223 84 6177/6178 email: richard@vocalis.com

1.8 Acknowledgements

The EAGLES project has been fundamentally conceived as a community undertaking. This breadth of participation is reflected in the following list of active members, technical authors and contributors whose corporate efforts have cumulatively resulted in the EAGLES handbook of standards and resources for spoken language systems.

As well as the chairman/host/rapporteur management team and the members of the Working Group mentioned earlier, the following individuals have made substantial contributions to the construction of this handbook:

Principal technical authors: F. Bimbot, L. Boves, G. Chollet, K. Choukri, E. den Os, C. Draxler, N. Fraser, D. Gibbon, P. Howell, L. Knohl, V. Kraft, H. Ney, R. van Bezooijen, V. van Heuven, D. van Leeuwen

Contributors: W. Barry, C. Benoît, D. van Bergem, J. Blauert, M. Cartier, P. Dalsgaard, C. Delogu, J. Esling, K. Fellbaum, A. Fourcin, M. Grice, V. Hazan, U. Jekosch, D. Johnston, H. Klaus, K. Kohler, L. Lamel, J. Llisterri, F. Neel, G. Pérennou, J. van Santen, H. Steeneken, A. Syrdal, I. Trancoso, J. Wells, B. Williams, J. Zeiliger

Editorial Support: D. Gibbon, I. Mertins, J. McNaught

Part I

Spoken language system and corpus design

2 System design

Many speech processing systems are commercially available and several applications have been successfully set up in different areas. In order to generalise their use in man-machine interaction applications a "predictive model of performance" of the technology to be used needs to be derived by application developers as a function of what they judge as relevant factors. This model of system performance, correlated to the technology specifications, has to be gauged against the ultimate target which is the application requirements and user needs. This chapter is based to a considerable extent on previous publications by Choukri et al. (1988) and Moore (Moore 1988, 1994a).

2.1 Introduction

In this chapter we do not discuss the available technical approaches nor argue the merits of each. We simply indicate what could be the needs and expectations of an application developer, should the solution be a spoken language or voice processing system.

For speech technologies, as for many technologies, each solution is a unique one that involves several types of expertise. A technology provider may come up with an integrated system (speech recogniser and/or synthesiser with ergonomics, oral dialogue, communication, software and hardware experts) that outperform others, although the pure assessment of single modules may show equivalent performance.

As a first recommendation we suggest that application developers request information from several technology providers to be sure that they are aware of the major international assessment methodologies and standards when available and applicable. Afterwards the application developer should request detailed proposals for the specific application to be set up. For the technology evaluation it is of paramount importance to account for the application characteristics as it is very hard to measure the quality of a speech recogniser or the comprehensibility of speech synthesis output in an absolute manner.

Spoken language systems are an appropriate combination of several modules including recognition of speech input, recognition of speaker identity (verification or identification), speech output generation and synthesis (including speech coding), and/or man-machine interaction management.

A simple use of a spoken language system consists of recognising speaker utterances, interpreting them with respect to the application, deriving a meaning (or a command), and providing consequent feedback to the user (maybe a speech prompt or a system action). This is illustrated in Figure 2.1 for a speech input/output dialogue system.

System design 31



Figure 2.1: Spoken dialogue system

In order to generalise the use of such systems in different man-machine interaction contexts, a predictive model of performance (Choukri et al. 1988) needs to be obtained as a function of different identified relevant factors. The definition of those factors has to lead to a set of parameters that can describe a speech processing system. This description has to express two opposed points of view: that of the technology provider (designer) and that of the application developer (buyer). The two points of view have to be distinguished.

Designers should give proofs of the performance of their systems with a measure of the impact of any change. So the first contribution of this chapter is related to the technology supplier point of view, and aims at providing detailed guidelines for the specification of speech processing systems in order to explicate the operational capabilities offered by the technology. This will allow the technology providers to depict the system performance in a comprehensive way to the application developers.

Buyers need comprehensive information about how each system or device will perform in the specific conditions of their application. So the second contribution of this chapter is related to the application developer's point of view, which aims at providing detailed guidelines on how to express the requirement of applications that incorporate speech processing systems in order to make explicit the application requirements that should be met by the operational capabilities of the technology. This will allow the application developers to express their needs in a comprehensive way to the technology 32 Part I: Spoken language system and corpus design

providers.

The technology specification is complex enough, and has to go beyond the sole numerical value of 99% accuracy usually announced by the equipment suppliers. This rate depends on numerous parameters. Some of them cannot be easily quantified (Pallett 1985; Choukri et al. 1988; Moore 1988). In order to focus on the most relevant parameters one needs to adopt a multi-dimensional characterisation of the speech processing system. This characterisation will be called the "system capability profile" (an expression first introduced by Moore).

The application requirement is also a complex phenomenon, too complex to be reflected only by a transaction success rate, and should also be depicted as a multi-dimensional characterisation. This will be referred to as the "application requirement profile".

The objective of this chapter is to list the major factors that would permit definition of the above mentioned multi-dimensional space and moreover a way to express a matching process between the two spaces. It consists of forms with keyword entries that relate to the different dimensions as seen from the points of view of the technology provider as capabilities and system features, and of the application developer as requirements. For each module we will provide guidance for the general terminology and specifications, and elaborate algorithmic aspects, software and hardware implementations, system integration and other features.

2.1.1 System capability profile versus application requirement profile

To compute the best match between requirements and capabilities implies that a capability profile can be compared to an application requirement profile. The capability profile indicates the available technology through commercial products as well as through pre-industrial laboratory prototypes (the last stage of the prototyping process). It exhibits what can be done. The application requirement profile indicates the technology needed to satisfactorily meet the user expectations. It expresses what should be done. Fitting capabilities against requirements can be illustrated as in Figure 2.2.¹

Of course the issue is more complex and many dimensions are needed to represent the fitting of capabilities against requirements. Although in many cases the relevant factors can be atomised as few major parameters which are called *dimensions of difficulty* (Moore 1994a). If we imagine this as a simple three dimensional space one may have the two representations shown in Figure 2.3.

The application developer should look for the best fit between his application requirements and the capabilities of available technologies. Obviously some of his requirements may not be satisfied. Optimisation criteria should

¹Courtesy of Roger Moore (Moore 1994a).



Figure 2.2: Capabilities versus requirements



Figure 2.3: Specifications versus capabilities

be considered in order to evaluate acceptable shortfalls in some dimensions against the improvements gained in others. The compromise he has to accept implies adjustments of the requirements, keeping in mind a bottom-line global performance.

Moore expressed this suitability of a system for a particular application with a diagram that is simplified in Figure 2.4.

2.1.2 Technical features versus system capabilities

The technical features are not meaningful or meaningless. The technical features of each system are of paramount importance to other technology designers or technical persons aware of the state of the art. However these do not reflect the dimensions of the profiles mentioned above: capabilities or requirements.

34 Part I: Spoken language system and corpus design



Figure 2.4: Technology specifications versus application requirements

2.1.3 System in operation versus laboratory prototype

The requirements expressed herein apply to both commercial products and laboratory prototypes. Of course a lab prototype may not be packaged as a commercial system should be, and thus its implementation requires the intervention of a member of the development team, although the prototype specifications and design should be done with later requirements in mind.

2.2 Speech recognition systems

The expression "speech recognition system" is meant here as the module that deals with speech input (acoustic wave forms) to deliver either a label or a corresponding command. Usually the designers focus on three major elements, which are the vocabulary (complexity, syntax, size), the environment (bandwidth, noise level, distortion type), speakers (stressed/relaxed, trained/untrained).

The major requirements relate to:

• vocabulary, speech and language modelling,

- training material (if needed), the data collection platform, pre-processing procedures,
- speaker dependency and speaking modes,
- environment conditions.

In the following we will describe most of the factors that may occur in the specification process of a speech recognition technology or the expectations of the users and thus of application developers.

A speech recogniser is based on some speech modelling using various paradigms. The best known are Dynamic Time Warping (DTW), Hidden Markov Modelling (HMM), and Artificial Neural Networks (ANNs). Most of the approaches distinguish two phases: A training phase and an exploitation phase. The first phase is devoted to learning speech characteristics from data:

- acoustic wave forms,
- phonetic/linguistic descriptions,
- specific features, etc.

The material needed for this phase is important and will be elaborated upon in the following sections. The second phase, related to exploitation, consists of the use of the trained system to recognise speech input. The key characteristics of the recogniser are described within the following sections.

2.2.1 Speaker dependency

The system may be designed to tolerate a large variety of speaker variabilities. In this case the system is speaker independent and has to deal with a large population of users, mainly from the general public.

Other systems may be tuned to the voice of a particular speaker (a single speaker) and thus the system is speaker-dependent. We may also have a system that is adapted to the voice of a particular set of speakers (multi-speaker system).

Some systems may also be adapted to be utilised only by trained speakers instead of the general public (untrained speakers). Others may be used by consistent or inconsistent speakers, skilled users/unskilled users, a specific "population", or the general public. Some may be used, for a given language, by native speakers or by speakers with foreign or heavy (regional) accents. In order to achieve speaker independency a training phase is usually needed and uses a specified type of speech corpus. The characteristics of the required speech data have to be described in terms of

- type of data
 - speech acoustic wave forms,
 - acoustic data with phonetic labelling,
 - acoustic data with the corresponding orthographic forms,

- 36 Part I: Spoken language system and corpus design
 - acoustic data with the corresponding phonetic transcription,
 - acoustic data with the corresponding recognition-units transcription, etc.)
 - size of data (how many hours/minutes of speech),
 - number of speakers and how they are selected
 - arbitrary, or
 - with respect to some characteristics such as
 - sex,
 - age or age ranges,
 - physical state, psychological,
 - experience,
 - attitude,
 - accent,
 - etc.
 - acquisition channels
 - single microphone,
 - set of microphones,
 - similar telephone hand set, or
 - as many handsets as possible,
 - environment conditions (noisy, quiet, all conditions, etc.) and
 - many other constraints derived from the operating condition.

2.2.2 Speaker-adapted systems

The system may incorporate a speaker adaptation procedure that allows the system to learn the current speaker characteristics and thus improve its performance during the interaction. At the beginning the system may be used in a degraded mode (either speaker-independent or speaker dependent trained on another speaker) and ending up as an optimised speaker-adapted system.

The adaptation may have to be done by the application developer in order to tune the system to his specific application. Usually two approaches are used:

1. A static adaptation process:

One has to start from an off-line recording of data and a training phase before using the system. The system references are adapted to the new speaker once and for all. The duration of this process is important: it can be real-time or even last for hours. The speech data needed can be acoustic data without any manual labelling or manual pre-processing, or it may have to be labelled (orthographic plus phonetic). The speech corpus may range from a few minutes of speech to a few hours.

2. A dynamic adaptation process:

The system learns the current speaker characteristics while the speaker is

using the system. This may be done by user request if errors occurred during the application, or the system may automatically take into consideration the speech data uttered by the present speaker.

If this procedure is available the application developer has to know how to use it and he needs the know-how to collect and process the required data.

2.2.3 Speaking aspects

2.2.3.1 Speaking mode

The first feature is the speaking mode: it can be isolated (the words are pronounced in isolation with pauses between two successive words), connected (usually used when spelling names or giving phone numbers digit by digit), or continuous (fluent speech).

2.2.3.2 Speaking fluency

With the three speaking modes, the speech input can be spontaneous or read (scripted speech for data entry by computer operators or a text dictated to a secretary from a manuscript document).

2.2.3.3 Speaking rate

It is obvious that the speech production rate varies from one speaker to another. The speaking rate depends on the exploitation conditions in particular due to stressful operating conditions such as adverse physical environments. The speaking rate can be slow, normal or fast. This may be measured by the statistical distribution of the average number of speech frames within a given set of sentences. If the performance result is obtained with a particular speaking rate that is not used during the exploitation it has to be specified. A tool may be required by the application developer to measure speaking rate.

2.2.3.4 Non-speech sounds

The users are likely to produce acoustic sounds that are not relevant to the application, such as cough, sneeze, clearing one's throat, lip smacks, clicks, etc. These extra-linguistic phenomena (or non linguistic phenomena) may be considered as part of the speech modelling (implementation of the rejection mode described below), or may be tackled at the linguistic level or other higher levels.

The application developer has to know if these phenomena are handled or not, and how to tune the system for that purpose if any intervention is needed. 38 Part I: Spoken language system and corpus design

2.2.4 Vocabulary aspects

2.2.4.1 Lexicon size

The number of words that are recognised by the system has to be given. This may consist of a small set of words (small vocabulary of about ten words), a medium-size set (from 10 to 100), a large set of words (from 100 to 1000) or very large (over 1000 words). The vocabulary may be seen as a single dictionary or divided into several sublexicons downloaded to the application considering some dialogue phases.

2.2.4.2 Speech training material

The training phase is a crucial stage and involves several types of data and acquisition conditions. One may distinguish the amount of data characterising the environment and characterising the acquisition channels, the amount of data per speaker that depends on the current application and the amount that is independent of the current application, the time between speech acquisition for training and system use, and so forth.

The data may consist of acoustic speech waveforms of isolated words, or orthographically and/or phonetically labelled speech sentences (arbitrary sentences, application dependent, phonetically balanced). The corpus size is also important in terms of the number of words/sentences and the number of repetitions. The data may include extra-linguistic phenomena that occur. The training phase can be achieved by the application developer or by the technology provider if it necessitates some in-house know-how and hints. If the training is not done once and for all it is important to point this out. The duration of the training process has also to be reported.

2.2.4.3 Speech data collection

As noted above, the training phase uses speech corpora that consist of a collection of recorded speech samples with corresponding labels. The technology provider and/or the application developer can select the list of words and set up a collection platform for this purpose. The speaker may have a sheet of paper with the list of the words/sentences to be read or he may be requested to repeat the sentences played back by the system, or he may be asked questions, with his answers recorded to produce the speech data.

This process can be undertaken with or without supervision. The problem to sort out is how to judge the quality of the corpora (in terms of acoustic, phonetic, and linguistic coverage) and how to label the data (the labels can be orthographic, phonetic, acoustic segments with end-point marks, etc.). Selection of appropriate acquisition conditions is a crucial matter and has to be done carefully (in respect of how and by whom it is produced). Data acquisition necessitates a platform with particular requirements regarding memory storage, CPU capabilities per word/sentence and speaker, and adequate user-friendly interfaces. These parameters are listed within Section 2.6.2.

2.2.4.4 Lexicon generation

The recognition process may use a set of words (global approach) or a set of subword units to identify the user utterances. If the system uses whole word models (global approach), these have to be learned beforehand. The vocabulary to recognise has to be recorded and used for training for each different application (fixed vocabulary). In the case of subword units (analytic approach), the speech units are learned once and for all and the vocabulary lexicon is generated as a concatenation of such units (flexible vocabulary).

In both cases the system may be optimised for a particular language or a class of languages. The multi-linguality aspect is of paramount importance in the era of open economic marketplaces.

FLEXIBLE VOCABULARY:

The subword units are acoustic units that may be based on linguistic or phonetic entities. The technology provider has to describe this as a system feature. The application developer has to know how to use the units to generate his own application vocabulary. During the recognition process a "parser" accomplishes the labelling of speech units and provides the lexicon entries.

For example, if the system deals with single words, the application developer has to know the task he will have to carry out and the skills needed for that purpose, such as:

- The application developer has to type in the orthographic forms of his lexicon (the phonetic transcriptions are automatically obtained via an orthographic-to-phonetic converter).
- The words have to be manually transcribed onto the recognition units.
- The words have to be transcribed and alternative pronunciations have to be introduced.

FIXED VOCABULARY:

In this case the system is task dependent. For each task there is a need to acquire specific and tuned corpora. The technology provider has to describe what kind of data is needed. This could be speech waveforms of isolated words, or labelled sentences, or labelled and phonetically balanced sentences, etc. He has to give the validity of the speech database in terms of the number of words, sentences, speakers, different acquisition conditions, etc. There may be a need for end-point detection (speech segmentation

40 Part I: Spoken language system and corpus design

and labelling) to be done manually or semi-automatically. There may be some differences in the database characteristics in order to come up with a speaker dependent or speaker independent system. If some extra-linguistic phenomena are likely to occur, the technology provider has to instruct the application developer how to take them into account.

The training phase can be achieved by the application developer or by the technology provider if it necessitates some in-house know-how and hints. Training is a time-consuming process and its duration has to be estimated by the technology developer.

For both fixed and flexible vocabulary approaches there are several possible procedures, which can be itemised as follows:

- The lexicon is based on subword units and has to be generated, for each new task, by the technology designer.
- The lexicon is based on subword units and may be simply generated, for each new task, by the application developer.
- The lexicon is based on whole word models, and a speech database has to be acquired, validated, and models trained, for each new task, by the technology designer.
- The lexicon is based on whole word models, and a speech database has to be acquired, validated and models trained, for each new task, by the application developer.
- The system is based on a hybrid approach where some lexicon entries are obtained via direct recording (fixed vocabulary such as digits, yes/no, where a list of such words has to be supplied) and others via subword unit concatenation (flexible vocabulary).

2.2.4.5 Speech modelling and syntactic rules

Some systems offer the possibility to define a particular syntax to recognise connected digits, sequences of words or to spot keywords within a sentence. This can be carried out at the algorithm level or at the modelling stage. The application developer should know what the options are and what the tools are that will allow him to use them (model generation or rule definition). For example, some systems offer a word spotting functionality through speech modelling and a syntactic rule formalism. Spotting words within a stream of speech is used to cope with para-linguistic factors such as hes-

itation (*Er yes* instead of *Yes*), polite styles (*Yes, please*, instead of *Yes*). This approach may be implemented using "syntactic" rules such as:

The tools may allow a model to be set up for fixed-length sequences of words (if the application expects 8 digits, then this piece of information should be used to increase the recognition rate) or this may be left to the user and the application may analyse the number of words and then manage the dialogue accordingly. This is a way to account for likely user responses.

2.2.4.6 Field speech data (in-situ recording)

It has been confirmed from many reported evaluation experiments that the speech uttered by the users during an exploitation session is more representative of the field conditions than the speech acquired during recording sessions. Such speech data may be stored, labelled, incorporated in the training database to account for the field characteristics, and can be used to provide new releases of upgraded speech dictionaries for the application developer. It is important to know whether this is possible or not and how to handle such material.

2.2.5 Branching factor (perplexity factor)

Most of the applications can be organised as a sequence of dialogue phases where each phase needs a specific lexicon. For example in a menu-driven dialogue system each menu offers a list of choices with the corresponding commands: a subset of recognition vocabulary that is active at one time. It is clear that if the active vocabulary is reduced to these commands the confusion between words will be reduced as well as the time devoted to the speech comparison, while increasing the recognition rate. For example if the system requests an input that should be YES or NO within a complete lexicon of 100 words then there is no need to compare the input to all the 100 words. Nevertheless one has to consider the alternatives, synonyms, and the possible extra-commands that may be pronounced by the user (such as No thanks, Yes please, Operator, Stop, that's all, ...). The branching factor is a "divide to conquer" approach (Kinsey 1994) that indicates that the technology supports vocabulary subsets. This may be achieved automatically if the dialogue manager or another integrated tool allows it or it may have to be done at an early stage either by the technology provider (e.g. HMM modelling of classes of words) or given by the application developer as a list of sublexica to activate each time the speech recogniser is on.

2.2.6 Rejection mode

It is clear from different field trials that the application users may be involved in other conversations and are likely to elicit words out of the expected lexicon such as extra-words, synonyms, and extra-linguistic sounds. The user may be using the system in noisy conditions and if the system is sensitive to the environment noise it will detect a word even if the user did not speak. In order to take such phenomena into account the system has to possess rejection capabilities.

42 Part I: Spoken language system and corpus design

Different approaches allow implementation of such functionality. The most common approach is a rejection model which is trained on speech data from words that do not belong to the application vocabulary (called *trash model*). In isolated word HMM systems, the trash model is also an HMM or several HMM models beside the models of the application vocabulary.

Another approach is usually based on thresholds about the speech recognition scores (likelihood, distance measures, etc.).

If the option of rejecting out-of-lexicon words or/and extra-linguistic phenomena is offered, then the technology provider has to explicate whether:

- The rejection capability is independent of the vocabulary (vocabulary and task independent).
- The rejection capability has to be developed with a detailed technical consideration of the lexicon of each application (involving some knowhow).
- The rejection capability has to be tuned for each task on the exploitation site by the application developer and/or by the technology provider.

2.2.7 Application vocabulary and confusion matrix

The technology provider performs different evaluation tests. These can be performed with laboratory databases as well as with field databases (speech recorded during exploitation). The confusion matrix permits a more detailed evaluation of the system capabilities than the usual recognition rate, with respect to a given database. It gives the confusion between words used to evaluate the performance. An example with some isolated French words is given in Table 2.1 for illustration (Choukri et al. 1988):

Table 2.1: Confusion matrix

	Classer	Placer	Annexer
Classer	66	22	5
Placer	19	59	1
Annexer	9	7	76

This matrix shows that in 22 cases (out of 93) the word "Classer" is recognised as "Placer". This matrix is very important because it can be taken into account to design an accurate recogniser through an adequate vocabulary selection where the most confusable series are withdrawn and replaced by synonyms or equivalent terms. It can also be considered at the application level to provide an optional candidate not supplied by the recogniser.

System design 43



Figure 2.5: Result presentation with "speaker selection" procedure

2.2.8 Speaker selection

Diagram 2.5 establishes the distribution of errors across speakers. It indicates the quality of the test database, and shows that there may be fewer than 2 errors for 99% of the speakers ("sheep speakers") and about 10 for some "goat speakers".

The two approaches described here (confusion matrix and speaker selection) may be influenced by the fact that building databases can be biased, and should be carried out carefully.

2.2.9 Language modelling

The language model is used to improve the performance. Different paradigms can be used: finite state syntax, external NLP module with a parsing grammar, an n-gram (unigram, bigram, trigram) model, etc.

The language modelling (see Chapter 7) may be tuned by the technology provider with respect to a particular variety of languages. An adaptation to a specific application can help to improve the results, and may or may not be necessary. This tuning may be done by the technology provider or by the application developer. A tool or a formalism that allows data to be entered may be provided and any data that has to be collected by the application developer should be indicated.

2.2.10 Channel adaptation/Environment adaptation

As described above the system may incorporate an adaptation procedure that upgrades the recogniser references in order to account better for current speaker characteristics. A similar procedure maybe set up in order to adapt the system features to the channel characteristics (a new microphone or a telephone line) or to the environment (room acoustics, background noise).
Requirements regarding this adaptation are similar to the ones described for the speaker adaptation process.

2.2.11 Task/Application adaptation

The system may need a particular tuning to suit a particular application or a class of applications. For example, generation of an adequate vocabulary and adaptation to the channel and/or background noise may be needed. Such tuning can be achieved through a "black box" procedure by the application developer, or may demand data acquisition and modelling that involves specific know-how from the technology designer.

2.2.12 Speech recognition and application interfaces

The speech recogniser output concerns the raw information returned to the application. As mentioned above this may be a word label (or a lexiconentry identifier), returned with other detailed data such as time-length of the speech signal, the energy level (which may indicate whether the user speaks loudly or not). The system may return the N best candidates: that is, the words recognised ranked with respect to their likelihood (probability scores or distance measure). Continuous speech recognisers may return a parsed sentence or a lattice of lexicon entries for analysis by a linguistic module. A word spotting system may return the number of recognised words and their respective labels with the time-frames of their occurrence.

For systems that use beeps to indicate the user's turn, the application developer may need to know whether the user speaks too close to the beep, does not speak, speaks loudly, speaks in much too noisy conditions, etc. Such information is relevant and may be processed at the application or dialogue stage.

The application developer may need to set up some of these parameters and should have access to them.

2.2.13 Speech input and speech signal acquisition

As noted above, a wide spread of information databases will lead to a widespread use of telecommunication systems to access remote databases. In the meantime, speech control of workstations, office and working environments (the desktop arena) opens new market sectors. It is believed that the wide spread of telephone handsets will permit adoption of speech technologies despite many limitations. The speech input can be either acquired through telephones or microphones.

Both media yield different performances and are suited to specific types of application.

2.2.13.1 Microphone interfaces

Using a microphone is an important means of improving performance, as it does not have the drawbacks caused by the technical limitations encountered with the public telephone network, such as bandwidth limitations – although degradation may occur due to the changes in microphone characteristics (electrical and acoustic).

The application developers have to be aware of the most important characteristics that are relevant to the speech signal representation and thus to the speech recogniser:

- channel and transducer characteristics such as:
 - frequency bandwidth,
 - distortions,
 - echo and echo delay.
- environment characteristics:
 - noise type (interferences, reverberations, background stationary noise),
 - signal-to-noise ratio.
- operating conditions:
 - accelerations and movements,
 - good positioning of the microphone: headmounted or handheld,
 - good positioning of the speaker regarding the microphone installation.
- other characteristics that may be important, such as the mechanical effects that may occur when using a press-to-talk microphone.

The application developer has to be aware that some users are not comfortable when talking to a machine and will not accept a headset microphone easily. Moreover it is obvious that if speech control is used it is not efficient to encumber the users with a handheld microphone. The use of a remote microphone attached to the monitor or to a stalk which can actively track the speaker can be suggested for real applications. In all cases the choice of a microphone is the responsibility of the application developer (who may consider other "human factors") but has to be clearly validated by the technology provider in order to ensure a high signal quality and a high signal-to-noise ratio to be passed to the speech recognition module.

In general the microphone to use during the exploitation phase has to be similar to the one used during the training phase. So the application developer should be instructed about its characteristics and the most "influential" factors in order to select an equivalent microphone.

Some technologies are well adapted to the use of microphone arrays, although this is still too expensive a solution for low cost and cost-effective

applications. In other cases the speech input is acquired through a telephone handset without the telephone line and environment. If such possibilities are offered they have to be clearly mentioned to the application developer. Some systems allow the use of other microphones provided the system is adapted to the new microphone characteristics through an adaptation procedure that modifies the speech references – trained with another microphone – or the speech input in order to meet the characteristics of the new conditions. In general, the systems have to be trained for each new speaker and for each new microphone.

2.2.13.2 Telephone interfaces

The telephone is becoming a new "computer terminal" and allows access to different services either using DTMF (Dual Tone Multi-Frequency, or touch tone), pulse detection or speech recognition. In almost all applications speech recognition has to be telephone-independent. The telephone channel includes the telephone handset, the private switch (PABX – Private Automatic Branch Exchange), the public network (PSTN – Public Switched Telephony Network) and the speech-based system interface with the public network (directly or through a PABX).

The use of telephones induces many phenomena that are not observed when the training is carried out with high quality microphones. Such phenomena include local acoustic echo, electrical echo, line noise, non-linearities, spectral distortions, etc. All these features are considered as non-linguistic sources of variability and have to be taken into account.

So the system has to process as its input signal the speech signal uttered by the caller and sent through the telephone handset, the public telephone network and the local interfaces. The system may have to provide the service, using the speech recogniser, to callers wherever the call originated. Hence the system has to support all types of telephone handsets and lines, unless the callers are a selected set of users who are asked to use specific handsets that are provided. In the general case the system has to accommodate telephone handsets with electret and carbon button microphones, cellular phones, etc. This is usually managed through the database collected and used for training. If the system is dedicated to a particular type of call (domestic, inter-city, long distance by satellite or undersea cable) then the technology provider should indicate whether he supplies the corresponding and appropriate speech models or whether he requires the application developer to collect such data. The system may also tackle the telephone line characteristics through an adaptation procedure and the application developer should be instructed about how to use it (see channel/environment adaptation section).

INTERFACE BETWEEN THE NETWORK AND THE SPEECH SYSTEM The connection between the public switch (PTT network) and the speech recognition system can be either analogue or digital. The application has to deal with telephone signalling as well as speech signals. Telephone signalling is analysed as part of the speech analysis module for systems that simultaneously manage speech signals, DTMF, and/or pulse detection.

ANALOGUE CONNECTION The interface between the system and the network is through the classical tip and ring wires. The signal is analogue and there is a need to convert it to a digital form. For this purpose one needs a telephone interface that answers the incoming phone calls (inbound calls) by an analysis of a loop-current drop event (go off-hook) and that manages the different telephone tones (dial tones, busy tone, etc.) with respect to the PTT regulations as the interface must fulfil the requirements for standard switching compliance (timing and control signal management).

Although the speech recogniser does not deal with telephone control signals, these have to be considered at the application level: to go off-hook, to know that the caller is still on-line, that he hangs up (goes on-hook), to detect an event related to the loop-current drop and to terminate signal processing, to process call transfer functions, to terminate a call, etc.

In some cases the systems host the telephone interface and the speech recogniser on the same board and speech processing functions consist of a set of programming functions of the same level as the telephone functions.

In many other cases the telephone interface is handled separately on another board and speech data is provided through specific expansion buses such as PEB, AEB, MVIP or SCSA described in Section 2.6.2. In that case speech recognition deals with speech frames forwarded by the telephone interface and does not care about the line signalling.

- ANALOGUE TO DIGITAL CONVERSION The speech waveform is converted to digital samples before being passed to the recogniser module. The converter characteristics are of paramount importance and have to be clearly identified. These are the sampling frequency (how many samples per second), and the coding rate (how many bits are used to represent each sample). The standard figures are 8000Hz and 8 bits but some systems may offer 8000Hz and 13bits. These parameters can be set up once and for all at the application level, but this has to be done in accordance with speech recognition specifications. If they can be configured then the application developer has to be aware of it.
- DIGITAL INTERFACES The telephone network is increasingly based on digital switches. The speech signal uttered by a caller is converted to a digital code and transmission between the public switches is done as a digital signal. If the link between the application and the local switch is through an analogue line then the signal is converted from digital to analogue. If the connection is digital then the switch provides a digital signal as 64 kbits/s of speech plus the telephone signalling. The connection can be a single ISDN connection (called S0), an equivalent of one line, or it can be an E1 (USA) or T2 (Europe) group of 24 or 30 lines.

The speech signal is acquired by a telephone interface and has to be passed to the speech recogniser through an expansion bus like the ones mentioned above (AEB, PEB, MVIP or SCSA, described in Section 2.6.2).

- 48 Part I: Spoken language system and corpus design
- NOISE CONDITIONS The telephone applications have to operate despite the tremendous amount of noise that comes with the calls. The collected speech includes other signals such as TV, radio, computer and printer noises, other speakers, car noise for cellular calls, etc.). The application developers have to know the best way to take all this into consideration.

Some systems tackle such noises through rejection capabilities and others through an adaptation of the signal-to-noise ratio. Some systems offer a gain control which can be set up automatically and dynamically, others offer a static gain parameter adapted to the general operating conditions. Instructions about how to optimise it have to be given.

Some systems that offer echo cancelling functions necessitate an adjustment of the input and output gains. Hence the speech prompt level has to be carefully adjusted in order to minimise the echo to allow efficient cut through or voice stop capabilities. Tools should be provided in order to set up such a level even if the prompts recording is carried out at a professional studio.

APPROVALS In order to connect equipment to the public network it has to fulfil the requirements for switching compliance and local regulations. These requirements include signalling as well as electrical and electromagnetic radiation. After the equipment has passed the tests, it is given a registration number that is requested by the local government agency before system deployment. This does not indicate any assessed performance but simply that the equipment connection to the network is authorised. In several countries the PTT approval concerns also the equipment that is behind a private switch (PABX), and in many others it concerns the application as a packaged system (software and hardware components). For example, if the application is implemented on a PC with a telephone interface board and a speech recogniser board then the whole package has to pass the regulation tests.

The application developer has to know the status of the system he is planning to deploy; otherwise he will have to take it out of the network.

2.2.14 Cut-through versus voice-stop (anticipation)

A major drawback of current speech technologies is the fact that when the system is prompting messages (outgoing) it is unable to listen. That is, speech recognition and speech prompting do not occur simultaneously. In most applications the caller has to wait for a system's instruction to be completed before speaking. Usually the users are instructed to "please wait for the beep before speaking".

Some technologies offer *cut-through* capabilities; these allow users to interrupt the system and anticipate on its messages. Speech recognition takes place even during the system prompts. This capability is highly appreciated by experienced users who may interrupt the system's outgoing speech any time. Moreover they may use a sequence of words instead of a single one. Of course it is necessary to carefully handle the dialogue structure in order to avoid interrupting the system because of noise or unexpected words. This capability has to be distinguished from the so called *voice stop* where the system detects some speech activity (energy level), stops prompting its present message, and switches to the recognition mode expecting the user to say something. Usually the system instructs the user to say "STOP" whenever he is ready to input his choice and then wait for the beep before speaking. In that case the beep is still there and speech recognition does not occur simultaneously to the speech playback.

The application developer may use these capabilities or may not (when he wants the users to listen to a complete message for security reasons). So the technology provider should indicate whether this is optional, mandatory, or not implemented. The distinction may be defined as follows:

VOICE STOP: In a first step the system hears, it needs a second step to understand. CUT-THROUGH: The system hears and understands simultaneously (single step).

2.2.15 Error measure presentation

Another important way to summarise the error distribution is to consider that from the dialogue point of view some errors are more severe than others. We usually distinguish the following types of error:

Table 2.2: Error presentation

Correct input	Input word is recognised.
	Input word is rejected (false rejection).
	Input word is omitted (not detected).
	A different word is recognised (substitution).
Incorrect input	Input is rejected.
	Input is accepted as a word (false acceptance).

2.2.16 Error handling

Different types of error may occur. Some of them may be handled at the application level, others may/should be handled at the speech recognition level or at least forwarded to the application.

2.2.16.1 Time out

What happens if the speech recogniser incorporates a speech detection activity and does not detect any speech signal during a period of time? This may be due to the user who does not speak or may be due to the speech detector. So this can be handled at the dialogue level as a time out parameter or it can be used by the recogniser to set up (or adapt) its thresholds of speech end-point detection.

2.2.16.2 Signal-to-noise ratio

The system may have some difficulties in extracting speech data from a continuum of speech with a low signal-to-noise ratio. If this is detected the application may ask the user to speak louder or use another handset. Some parameters may need a particular tuning (e.g. energy thresholds) to avoid significant performance degradation with a change of microphone or telecommunication channel.

For example if the system has been optimised for adverse and severe conditions and is being used in benign or favourable conditions (or vice versa) this may lead to significant degradation of performance and the application developer should be informed. He may require some figures about the performance obtained with different signal-to-noise ratio levels.

2.2.17 Response time

The response time is related to the integrated system and indicates the time needed to recognise a word/sentence and issue the corresponding action. It is computed as the time between the end of the spoken word and the beginning of the action. In general the speech recognition process consists of two separate modules: a real-time one concerns the signal acquisition and analysis and a second module – non-real time – for the speech comparison with references. Usually two implementations are used:

- 1. The first approach starts the recognition process (signal analysis, parameter computation, comparisons, etc.) as soon as the speech detector indicates that the speaker is uttering something. It stops when the end-point detection declares that the end of a word is detected and a valid word is recognised. The speech recognition runs synchronously with speech input acquisition. The response time depends on the time needed for word detection and the comparisons.
- 2. The second approach consists of initiating the speech recognition when the system declares that it detects both word end-points (beginning and ending times).

The application developer has to know what his system should tolerate as recognition response time and should take into consideration the different processes that may run on the same machine.

The application developer has to know to manage his CPU load when using a multi-channel system and should require a uniform and coherent response time on each channel. The technology provider should guarantee maximum response time in the worst conditions. This of course depends on the application: a response time may be acceptable for single isolated words and may not be acceptable for recognition of a sequence of words (connected digits or words).

2.3 Speaker verification/identification

2.3.1 Speaker verification versus speaker identification

Speaker verification aims to verify the identity of the speaker through a comparison of some samples of his speech with the references of the speaker he claims to be.

Speaker identification aims to identify a speaker who belongs to a group of users through a sample of his speech.

One should distinguish the use of speech to identify/verify the speaker identity on the one hand from the recognition of the utterances he produces (password, PIN) to check his identity explicitly in a database on the other.

2.3.2 Speaker characteristics verification or identification

Other tasks of classification of speech patterns according to specified characteristics of the speaker are proposed by the technology providers:

- To decide whether the speaker is male or female, (sex identification),
- to estimate an age range of the speaker: child, young, older (age identification),
- to detect the speaker's current emotional state using speech samples (mood state identification).
- to detect any pathologies using speech samples (health state identification)

The application developer may think of other characteristics which could be available for use within real applications. Such tasks are not detailed within the next sections although many application requirements apply to these tasks.

2.3.3 Error measure

2.3.3.1 Speaker verification errors

The different errors are summarised in Table 2.3.

Table 2.3: False acceptance versus false rejection

	ACCEPTED	REJECTED
The speaker is	The system accepts	The system rejects him
user U	him as speaker U	(false rejection)
The speaker is an	The system accepts	The system rejects him
impostor who claims	him as speaker U	
to be user U	(false acceptance)	



Figure 2.6: False acceptance versus false rejection

The application developer may have access to the different parameters that determine the system performance. These parameters may be directly correlated to given threshold or confidence levels. For example, he may accommodate the thresholds in order to obtain an equal error of false acceptance and false rejection as well as a confidence area as depicted in Figure 2.6. The technology provider should indicate how to handle such parameters. He also should indicate whether such parameters are speaker independent or speaker dependent. If these are speaker-dependent parameters then he should know how to obtain and optimise them.

2.3.3.2 Speaker identification error

The speaker identification process depends on the size of the population, and the criteria indicated above for the speaker verification process should account for that. The error measures should also consider the confusion that may occur between two different speakers (the substitution possibility of identifying user E instead of the present user U). A particular summary is given in Table 2.4.

Substitution errors are more severe than the others because unauthorised speakers or impostors may thus gain access to confidential data. As the prime motivation for integrating speaker identification procedures is to achieve more reliable personal identification in a convenient manner this has to be used with other techniques.

System design 53

	Accepted	Substituted	Rejected
The speaker is the user U	The system accepts him as speaker U	The system accepts him as user E (substitution)	The system rejects him (false rejection)
The speaker is an impostor	The system accepts him as speaker U (false acceptance)	The system accepts him as user E (substitution)	The system rejects him

Table 2.4: Different types of error

In practical application one imagines that the users are motivated and hence are very cooperative. Meanwhile the impostors are unknown speakers and there is no way to collect data to prepare a rejection model based on an "impostor model".

The application developer has also to know how to calibrate the different thresholds to obtain the best compromise between false acceptance, false rejection and substitution errors.

In some particular applications dealing with speaker verification, a confusion matrix may be requested in order to allow a pre-selection of several candidates in a first phase and then consider a second process with a small subset of speakers.

2.3.4 Training

The following description applies to the speaker verification as well as the speaker identification process. The system may be based on the speaker uttering a sentence or a sequence of words to give some samples of his speech. The comparison during the exploitation phase uses a reference dictionary obtained during the training phase.

The training phase may be off-line or on-line and carried out at:

- The technology provider site,
- the application developer site,
- the customer site.

The training phase may be carried out off-line using a particular platform, or on-line while the application is operating. The application developer has to know whether he can achieve the training himself (or the end-user can do it) or he will have to deliver the data to the technology provider who will provide the speech models.

The training material can be specified by the technology provider as a list of phonetically balanced sentences, well chosen sequences of words, or data selected with respect to some particular criteria (e.g. phonetic coverage of

the language). In some cases this material has to be collected and modelled by the application developer. In some other cases it is automatically done during a training session that is seen as a black box procedure. In all cases the technology provider should indicate the size and characteristics of the speech database needed to achieve the required performance.

The system documentation should also indicate the kind of know-how necessary to best exploit the technology if the training is accomplished by the application developer. This may be a list of appropriate phonetically balanced sentences per language if this is required, a tool to generate a minimal set of sentences or words, a selected list of words, etc.

If the training is achieved off-line using a database that has to be recorded beforehand then the application developer has to know what intervention is necessary to obtain a usable corpus. These can be speech segmentation, speech labelling using phonetic labels, orthographic transcriptions, etc. Consequently the application developer should request an adequate development platform with adapted tools such as a speech recording and analysis environment.

2.3.5 Exploitation

As mentioned above, this handbook contains a chapter dedicated to speaker verification and speaker identification (Chapter 11) giving details about the technical approaches. A particular section of that chapter is related to the "levels of text dependence" with a focus on the text-dependent versus the text-independent approach and a mixed strategy.

2.3.5.1 Text dependent approach

The system may use a text dependent approach: the user has to pronounce a specific sentence (or set of words/sentences) expected by the system. The application developer has to know how appropriate they are to his application. If they are not, he has to know how they have been selected. It is convenient that such a set of sentences can be changed from time to time; therefore he has to know how to choose new ones and what kind of training is needed (A new training phase has to be carried out, slight and minor adaptation is recommended, nothing to change!).

2.3.5.2 Text independent approach

The system may use a text independent approach: the system does not know what the user will utter. Typically the system asks questions (personal ones such as *How old are you*, *How many children do you have*, etc.). The system uses the uttered answers to identify/verify the speaker and may use the answers (a speech recognition is then activated to recognise what has been said) to double check in a database for consistency. The application developer has to know whether some specific tuning is needed. There are some hybrid systems which use a pre-defined list of words/sentences from which they choose in an arbitrary way the word/sentence to be spoken by the user. The system can also mix the two approaches starting with a text independent system and ending up with a text dependent one.

2.3.5.3 Speech quantity and quality

For the training process an amount of speech is needed to achieve a given performance. If the data collection is carried out by the application developer he has to know the size of the corpus to be collected.

The technology provider has to indicate the identification/verification performance correlated to the speech input duration which is shown in Figure 2.7.



Speech input duration

Figure 2.7: Error rate related to speech input duration

The speech duration may be used for a pre-selection of a subset of speakers (using for example a short sentence of less than 2s) and then request another sentence to identify the speaker.

The quality of speech is one of the major influencing factors and mostly depends on the acquisition device and conditions (microphone, telephone, bandwidth, acoustic environment, etc.). Some of these factors have been pointed out in the previous section about "Speech recognition systems" and will be elaborated upon in Chapter 10.

The technology provider may need to collect speech data regularly for training purposes because of the variation of speech and speaker characteristics in the course of time. This has to be clearly stated and the application developer should know how to accomplish this task and how frequent it is.

2.3.5.4 Adding / Removing speakers

Some systems carry out training based on a discriminant principle that involves a mutual information process. The optimisation is done so as to account for all speakers. In this case, adding a new speaker or removing one from the list of speakers to be identified requires a complete new training session. This may be done by the application developer using a black box procedure or he may use some hints and know-how that should be clearly indicated by the technology developer (e.g. adjustment of thresholds). An intermediate solution may be offered to allow adding or removing speakers without proceeding to a new training. The performance degradation should be clearly estimated.

If the training process does not involve this mutual information then the application developer has to know how to modify his dictionary to add or remove a speaker.

2.4 Speech synthesis

System feedback is mostly given by speech prompts played back to the user (production of audio signals equivalent to human speech). Speech output is the only possibility one has in telephone-based applications. (For a detailed account on speech synthesis see Chapter 12.)

The system's response generation (selection of the appropriate message) is achieved by the dialogue module and will not be described here. The system's responses through speech outputs are of many types which can be itemised as follows:

- The speech messages are recorded and stored, and optionally compressed, then played back to the user (like a computer-based digital recorder).
- Elementary speech messages are recorded, stored, and concatenated to produce the system's messages (canned speech).
- The speech messages are generated through a text-to-speech synthesiser (TTS).

2.4.1 Speech recording, storage, and playback

The speech is simply converted from an analogue signal to a digital one and stored on computer using a pre-defined format. Under the application control a speech file is selected and played back to the user (e.g. converted from digital codes to an analogue waveform).

The speech acquisition may be done through an analogue telephone line, a digital line or a local microphone. The application developer has to know whether he can record such prompts and what kind of tools are delivered for this purpose.

The converter uses a sampling frequency which is usually correlated to the telephone bandwidth (8 kHz), although multimedia applications use 11 kHz

or $16 \,\mathrm{kHz}$. Some technologies incorporate both a telephone interface and a microphone.

Different coding techniques are used, in particular PCM (Pulse Code Modulation) and ADPCM (Adaptive Differential PCM). The most common PCM method consists of sampling speech data at a rate of 8000 samples/second, leading to a 64 kbits per second of speech (one byte per sample). There are two types of sample coding called μ -law (very popular in the USA) and A-law (popular in Europe). So the application developer has to know that this requires 64 kbits per second of speech (64 kbps), which he may have to play back to the users. This technique is standardised by the CCITT under Recommendation G711.

ADPCM proceeds by an encoding of the difference between two adjacent samples, leading to a 32 kbits per second of speech. This has been standardised by the CCITT under Recommendation G721.

There are other coding techniques that allow compression of speech with rates below 32 kbps, like CVSD (Continuously Variable Slope Delay) at a rate of 24 kbits/s, an Adaptive Differential PCM (ADPCM) at a rate of 16 kbits/s, subband coders with a rate of 24 to 16 kbps, etc.

The speech data may as well be compressed at a low bit-rate before storage. Of course there is a quality degradation the application developer should take into account. The compression algorithms allow reduction of the speech rate to 9.6 kbps, 7.2 kbps or 4.8 kbps. These compression techniques may be of interest to some application developers who need to reduce the storage capacity of their system and thus the technology providers should inform them about the available techniques.

For example, a subtle combination of different compression rates may be requested in voice mail applications. The application developer may want to store the new and recent messages using the highest available quality and archive the obsolete ones using a low-bit rate coding. If this combination of techniques is available, the technology provider has to provide instructions about how to exploit it.

Another crucial problem that may occur is the misinterpretation of messages when they are mixed with music and DTMF sounds. The application developer has to know how to handle this problem (modify the message or decrease the music segment by some octaves).

In order to achieve the process of recording, storing and playing back the system prompts there is a balance to look for depending on various parameters such as:

- achievement of a real-time process (what processor, how many MIPS),
- to obtain the best playback quality,
- disk and memory storage (what kind of memory, capacity),
- bandwidth requirements.

Such parameters have to be clearly stated by the technology provider.

2.4.2 Canned speech

In some cases the number of sentences that may be played back is too large to be recorded. For example if the application deals with flight services then the system should be able to give information about flight numbers such as: "Flight A F Nine Three Zero One from PARIS will land at ELEVEN HUNDRED TWENTY FIVE." It is obvious that the storage capacity and the recording effort necessary to obtain all the possible combinations of sentences are important. So the approach consists of recording independent speech segments such as: "Flight", "from", "will land at", names of all the quoted cities, the digits, and the alphabet. The sentence to play back is obtained by linking appropriate segments together through a substitution of individually stored words in the information slots of the carrier sentence:

"Flight" {alphabet⁺} {digits⁺} "from" {city} "will land at" {hours}{minutes} ... This is usual in applications like schedules, bank balances, etc.

The scenario to perform consists of listing all the sentences that may occur. The application developer should know whether the technology provider supplies any tool that generates all the possible sentences (generation of a written version of the sentences to check). He should be able to select the words to be individually stored (the variable parts) from the carrier sentences (the fixed ones). Afterwards one has to define the recombination rules to account for the language specific characteristics (assimilation of adjacent words, coarticulation effects, etc.). For example in some languages like French there is a particular liaison that imposes contextual rules and exceptions: the number "21" is pronounced /vẽ te ẽ/ and cannot be produced by simply concatenating 20 and 1 (/vẽ e ẽ/). The corresponding rules should be delivered by the technology provider or implemented by the application developer using a suitable development environment.

2.4.3 Text-to-speech synthesis

As pointed out above speech synthesis assessment methodologies are at a less advanced stage than other technologies. Chapter 12 recommends several approaches to tackle the most important issues.

Text-to-speech (TTS) is the possibility to convert textual information into speech output. This technique is used by systems that handle huge pieces of information or information that varies frequently (products in a catalogue, daily press releases, access to e-mail messages, etc.). Another type of speech synthesis is based on concept-to-speech (CTS) and may be integrated into dialogue systems (cf. Chapter 12 for a brief description).

A TTS system basically involves the following modules:

- A linguistic module to manage the graphemic forms of texts,
- a phonetic module,
- an acoustic module.

It is obvious that an end-user and an application developer are mainly interested in the quality and performance of the speech synthesiser as a whole. The synthesiser is to be considered as a *black box*, although some applications may need particular tuning of some modules to account for a specific feature of an application. For example a reverse-directory application may need to pronounce proper names, including foreign ones; other applications may need to pronounce particular abbreviations or acronyms (military or scientific domains). For this purpose the application developer has to know the different components of the speech synthesiser (*glass box analysis*). He should know how to pin-point the module that causes a problem and whether he has any control over each component to correct the problem or to balance its effect.

2.4.3.1 Linguistic part

The linguistic part usually incorporates a pre-processing stage to deal with initial input data which is not "standard text". The data may include non-alphabetic symbols (parentheses, punctuation marks), abbreviations, acronyms, and exception entries. This pre-processing has to replace the symbols by the corresponding text, to expand abbreviations, to complete the correct pronunciation of acronyms, to state whether a string of digits is a number or a sequence of single digits, to correct the orthographic and syntactic mistakes. The application developer has to know whether he has any control over this submodule and at what level. To illustrate this he may want to get the appropriate French pronunciation of the acronyms and the corresponding (in this case French) pronunciation: CNET /knEt/, CCITT /se: se: i: te: te:/.

He may also be forced to replace systematically these acronyms in the input text whenever they have to be pronounced. The application developer has to know how to handle this.

Of course the application developer should not deal with the assessment of the linguistic module but only get the best tools/hints to improve it (or to tune it) if he selects the speech synthesiser for its global performance as a black box.

The linguistic part is also in charge of converting the pre-processed orthographic text to its phonetic transcription (or generally any abstract elements representing the speech sounds). This may use special rules as well as phonetic dictionaries. At that stage, words requiring special pronunciations are considered. Syntactic and lexical analysis are carried out to assign "lexical stress" and a "prosodic pattern" that will give synthetic speech its naturalness. This may be dependent on the application and has to be done by the application developer.

The output of this part is the decomposition of words into morphemes

and then into phonetic symbols with prosodic patterns (syntactic/lexical accent).

As it is detailed in Chapter 12, there is no standard and agreed-upon methodology for the assessment of the linguistic module of a TTS. Our intent here is to define the different tasks that a speech synthesis technology should accomplish, and the ones that will be needed by the application developer. These may be carried out automatically but for better tuning they may need some interventions as illustrated above. The underlying processes may or may not be offered by the technology provider. It could be possible to acquire some modules from another supplier and incorporate them with the synthesiser adopted, though this may not be easy. All these possibilities should be carefully taken into account. If the application developer has the intention or the need to modify subparts of this module he should know the format and any tool for editing the rule/dictionary component to be modified.

2.4.3.2 Phonetic part

This module concerns the construction of segmental synthesis using the concatenation of pre-recorded segment labels. It uses the broad phonetic output of the linguistic module to produce a set of the basic building blocks (which are diphones, triphones, individual words if any, etc.) that will be concatenated. This is based on an inventory of the labels of speech segments that are associated with "phonetic" sequences.

This module outputs a complete sequence of such "building blocks" with the appropriate prosodic markers (stress markers, accent position, etc.).

For example in a class of applications one may need to present a word with a particular focus to capture the attention of the end-user (listener) (e.g. departure time for a travel information system). This "accent placement process/rule" may or may not be offered by the system and may or may not be easily accessible to the application developer.

The application developer has to know who provides the necessary inventory of pre-recorded segments and how to create new inventories in order to personalise the voice output.

At this level one may need to incorporate a natural intonation contour and/or a duration model. This is usually done automatically. For special purposes the application developer may need to handle this manually and the technology provider has to inform him if this is possible and how to achieve it.

2.4.3.3 Acoustic module

The acoustic module aims at the generation of synthetic speech using prerecorded segments, extracted from acoustic speech data, using a voice coding algorithm. The technology provider may deliver a single output voice, male or female. He may also provide a tool to generate personalised voices (the aim here may be to have a company-tailored voice).

For all these modules, the application developer needs to know if human intervention is necessary to obtain satisfactory speech output.

The acoustic module may be based on a non-segmental approach, but the requirements indicated here remain.

2.4.3.4 Quality measures

The chapter devoted to speech synthesis assessment (Chapter 12) points out the different measures that should be conducted by the technology provider and demanded by the application developer. Such measures have to be conducted with the proposed factors and scales in mind. We may quote: naturalness, acceptability, intelligibility, listening effort, pleasantness, comprehension, and so forth. The evaluation should be application-specific or at least mention the way it is done. A TTS may be acceptable for the 1000 most frequent words/sentences but not the ones to be synthesised by the application under development. There is also an important item to consider: to measure the intelligibility of phonemes, words, and sentences, as some applications require understanding of sentences while others demand understanding of keywords (names, dates, digits), including some minimal pairs with no dialogue contextualisation.

2.4.4 Multi-linguality

The application developer also has to know which modules are language dependent and what the available languages are.

It has to be mentioned that some technology providers propose a "development environment" to facilitate the inclusion of new languages. The tools may be tuned to specific languages, may demand particular skills, and it may be time consuming to add a new language.

2.5 Interactive voice systems

An interactive voice system incorporates a dialogue module in charge of managing the interaction between the user and the system. The general case of interaction involves different media as the system's output and different modalities as the system's input. For telephony applications the input consists of speech input, DTMF, and/or pulse detection.

2.5.1 Menu-driven systems

The interaction is mostly structured as menu-driven sequences with a graph or finite state automaton model of the interaction. The transitions between the states of the automaton consist of a "single" command or action at a

time which can be a word/sentence recognition at a time or other application specific actions such as delays or noise level. Some systems offer a more sophisticated menu-driven dialogue where several actions are combined in order to proceed rapidly (combination of words, connected words/speech, etc.).

If such an approach is adopted by the technology provider then the application developer has to know how to implement his own application. Usually an application generator is supplied and allows rapid set up and evaluation of applications. Some integrate different ergonomic principles (e.g. management of time outs) and incorporate an error-recovery strategy. Appropriate information should be delivered for that purpose.

2.5.2 Sophisticated dialogue manager

Some systems offer a sophisticated dialogue manager based on an understanding procedure and different modules of the dialogue manager such as a user profile (to tune the dialogue to each speaker), a task model (in order to allow rapid customisation of the dialogue manager to suit a specific task), a user-belief module (to account for the expectations of the user at a given point in the dialogue). The objective of such dialogue manager is to parse and identify the application relevant parameters and to instantiate a semantic structure that matches the user query before accessing the application database.

2.5.3 Interaction with an information database

Spoken dialogue systems are useful for accessing data from telephones. The data may be local or remote. The implementation of the application should account for this, as accessing databases (particularly if they are remote) has an effect on system response time and also on the cooperativeness of the application: there is no need to access the database if in the previous call an identical query showed that the information requested is not available. The application developer has to know how to take into consideration the peculiarities of his application; for example, it may concern information retrieval from databases with no reasoning involved, or interactive problem solving with complex reasoning.

In designing the application, depending on the technology provided, some data may be available on the local system and should be downloaded from the remote-host in order to minimise the number of accesses to the host; others data may be remote.

2.5.4 Error recovery strategy

Speech recognition is not 100% reliable and some errors may occur. The dialogue manager may detect inconsistencies and focus the interaction within

a specific strategy. This strategy can be a switch between an open dialogue to a restricted one devoted to a yes/no question.

2.5.5 Routing to an operator

The dialogue manager may detect that the speech recogniser drastically fails with the current speaker and then may instruct the application to route him to a human operator (if available) or ask him to call another phone number to get a human operator.

2.5.6 DTMF, pulse detection, and voice recognition

The application may need to accommodate speech input, DTMF input and pulse detection. The application developer has to know how the system can accommodate such inputs simultaneously or has to choose once and for all at the beginning of each transaction.

2.6 System platform

The system that will host the application may consist, for example, of a PC platform, a SPARCstation, or a proprietary system like a VME-based system. The application developer has to be informed about the different platforms available. To this hardware box one should add some dedicated software and hardware.

2.6.1 Software aspects

2.6.1.1 Operating systems

The operating system of the platform may be a particular release of UNIX, OS/2, DOS, Windows, Windows-NT, or an in-house operating system. The application developer has to be aware of the operating system, as implementing applications will heavily depend on the operating system. For example, if the application has to handle more than one line per system then there is resource sharing, which is more easy to manage with a multi-task operating system than with a single task system like DOS. If the available technology runs under DOS then a spooling mechanism is to be used and should be clearly documented. Of course if the application developer is supplied with application development tools that make the operating system transparent for him then there is no need for that.

2.6.1.2 Drivers

If the adopted speech technology uses particular hardware for which drivers are provided, then the application developer may need to integrate them within his own software development environment. He has to know how to do that.

2.6.1.3 Application programming interfaces (APIs)

High level application programming interfaces may be provided and thus the application developer may use them directly with no calls to the drivers. Such application programming interfaces may be coherent with the best known ones such as TAPI, TSAPI or the planned SCSA TAO (SCSA Telephony Application Object). The aim of such APIs is to simplify the development of telephony applications independently of the technology providers and the hardware.

The application developer should know whether the technology he is buying incorporates such APIs as this will allow him to port his application easily to a new technology or a platform if required.

2.6.1.4 Application generators

Implementing voice-based applications may be achieved through the use of APIs as mentioned above or through user-friendly easy-to-use software called an *application generator*. The first strategy involves a programming task to be carried out by an expert who has to avoid many problems related to the integration of telecommunication devices and computing facilities. The use of an application generator permits building up of applications using a high level script language or very popular Graphical User Interfaces (GUIs). In both cases no specific know-how about hardware drivers and so on is required.

High level script languages allow the developer to write his application using a "natural like" language while graphically based languages allow the user to draw his application as a flow chart using the icons provided. The script functions or the icons are pre-programmed functions.

In both cases the application developer has to know what the preprogrammed functions are and - if expansion is provided for - how he can add a customised function of his own. Some generators allow functions to be written using a programming language (like C) and incorporate them together with the other functions.

Some application generators offer the possibility of generating either C-code or a script language. This may allow experts to customise the application better. The application developer should be informed about this.

2.6.2 Hardware aspects

The hardware aspects concern the complete integrated system delivered to the customer. We will distinguish several components that can be supplied by the technology provider or by the application developer. This chapter aims at decomposing such a system into its basic material components: platforms, speech processing boards, speech input/output interfaces, etc.

2.6.2.1 Platforms

The platform is the "black box" that will be installed at the customer site. It can be a PC – or compatible – or a proprietary system.

The major requirements are related to its capabilities in terms of CPU (386, 486, Pentium, Power-PC, Motorola), memory (RAM, hard disk), data transfer rate through the PC-bus (ISA, EISA, PCI) or through dedicated buses, data transfer rate from the memory cache to the disk (when writing files), the capacity in ampères required to power the expansion boards. Dedicated boards, with DSPs, will be used within the free slots available on the platform. These may be half/full slots of the backplane.

The application developer has to know what the hardware configuration is that will respond to his needs and then state the requirements as above. He has to know if there is any means to use more than one platform using a LAN.

2.6.2.2 Speech processing boards

For speech processing specific boards (a dedicated board or off-the-shelf from Dialogic, Rhetorex, LSI, NMS, other vendors) or local CPU capabilities may be used. The application developer has to know how to install and configure the boards. He also has to know the capabilities offered by the board with respect to his application. For example if an application has to recognise 10 words then the developer may use a single speech processing board to process two calls simultaneously. So he has to know about the number of simultaneous sessions/calls that can be handled in real-time (how many telephone lines if the system is telephone-network oriented). In some applications this depends on the language and thus has to be taken into account (a TTS board may handle 3 calls for Spanish synthesis but only one for French).

The technology provider may also offer different boards with multi-channel configurations (Board A = 2 recognisers and Board B = 4 recognisers). The application developer has to know whether he can plug in either of those and still run his application.

There are also hardware constraints about the number of free slots, the power and memory requirements, etc. that are needed.

2.6.2.3 Speech input/output interfaces

If the application is used within a desktop application, the speech input/output may use a sound board with an integrated microphone. If it is used within a telecommunication application then there is a need for an interface to the PTT network. This is provided by many vendors. As for the speech processing boards the application developer has to know what the requirements regarding his input/output interfaces are.

2.6.2.4 Connectivity

In many configurations one needs at least two boards: one to deal with telephone signalling (telephone interface) and a second one that implements the speech processing. The two boards use a particular bus to exchange speech data. The objective of such bus is to allow interaction between different boards implementing different applications from different technology providers on the same platform in an open environment. These are hardware and software implementations. The best known ones are:

- PEB (PULSE CODED MODULATION EXPANSION BUS) which is seen as an internal switching matrix capable of routing any time slot to an adequate audio port of the speech recogniser.
- MVIP (MULTI-VENDOR INTEGRATION PROTOCOL) is a multiplexed digital telephony highway for use within one computer chassis. It provides standard connection for digital telephone traffic between individual circuit boards. It supports telephone circuit-switching under direct computer control, using digital switch elements distributed amongst circuit boards in a standard computer. MVIP software standards allow system integrators to combine MVIP-compatible products from different vendors. The communication technologies that are supported include call management, voice store and forward, speech recognition, text-to-speech, Fax, data communications, and digital circuit-switching. The objective of an MVIP bus is to carry telephone traffic. It allows the interface to the telephone network to be separate from voice processing resources so the telephone interface may be obtained from one vendor while the voice processing resources are obtained from others. A single MVIP bus has a capacity of 256 full-duplex telephone channels. (Mitel, MT90810 Manual).
- SCSA (SIGNAL COMPUTING SYSTEM ARCHITECTURE (SCSA): According to Dialogic statement the SCSA represents the next generation of call processing architecture that opens up a new means of delivering and communicating information. SCSA is a comprehensive multi-layered hardware and software architecture for building call processing systems with multiple technologies and standard interfaces. So the objective is to provide standards that allow portability, scallability and interoperability with different developers' applications. An SCSA bus has a capacity of 1024 time slots (for a PC) on its bus called Signal Computing Bus.

The availability of such connections on the technology provided (hardware as well as APIs) allows easy portability of the application if this is anticipated.

2.6.2.5 Real-time aspects

As mentioned above, the application developer has to know how to manage his CPU load when using a multi-channel system and should require a uniform and coherent response time on each channel. The technology provider should guarantee maximum response time in the worst conditions. The real-time aspect is related to a complete application and should be estimated with all the lines on. For example if the system prompts a beep before starting speech recognition the application developer has to compute the delay: the beep prompt plus the time needed to start recognition. This time is crucial as people may speak before the beep, which leads to a gap error.

2.6.3 Planning for expansion

The system to be deployed may need to be expanded after a short period of time (to move from 8 lines to 16 or from few analogue lines to a digital E1 or T1 interface); so the application developer should be aware of what he should do for that purpose and take into account the cost of such a process when defining his requirements.

2.7 System simulation and prototyping

The application developer may need to simulate or prototype his application to identify the major phenomena that may occur, experimenting with a small set of users. Rapid prototyping may be done if an application generator is provided or an adapted application skeleton is available. For complex applications one may use the Wizard of Oz (WOZ) approach that is described in Chapter 13 of this handbook. The application developer should know if such tools are available and how suitable they are for his application.

2.8 Host–computer interfaces

In many applications, the IVR has to be connected to a host–computer to get and/or store appropriate data. The IVR is usually seen as a terminal or group of terminals, and each user is considered as a terminal session, on the network.

So there is a need for proper or standard communication protocols. Usually technology providers use host-computer emulation packages (like IBM3174, X25, ANSI-ASYNC, DEC-VT100, TCP/IP, Novell-Netware, etc.). Recently other interfaces were introduced to support the integration of telephone resources with computer-based information systems (see section on CTI below). The application developer should know if such connections are available and how suitable they are for his application.

2.9 Computer telephony integration

Computer telephony integration aims at the integration of computer-based systems, networks, telecommunication systems, PABX, and voice processing systems.

In order to minimise the time needed to develop applications or to port them to different computers or PABXs, there is a need for standards. It is more efficient to develop applications within a familiar development environment than learning a specific programming language and functions for each application.

The first initiative to develop such APIs originated from NOVELL/AT&T in 1993 through a TELEPHONY SERVICES API (TSAPI). Many PABX manufacturers supported it (ALCATEL, AT&T, FUJITSU, MITEL, etc.). The objectives were to allow Novell-Netware users to access the PABX functions from their workstations.

The second initiative originated from MICROSOFT, and is called TELE-PHONY API (TAPI, or MS-TAPI for MICROSOFT-TAPI). Its objective is to allow each PC user to operate a standard telephone from any Windowsbased application (dialling, voice-mail prompts on the PC screen, etc.). Windows'95 incorporates TAPI functions and a development kit.

Many companies offer applications or hardware platforms with TAPI or TSAPI functions (Dialogic, Rethorex).

So the application developer should know what kind of API he gets with the technology he is buying. The ones mentioned above constitute a comprehensive set of interfaces that will be widely available soon.

2.10 Multi-lingual aspects

The application developer may need to run similar applications in different languages. He has to know whether this is possible within the technology he is acquiring. He also has to check what kind of adaptation is needed when moving from one language to another. For example if the voice system uses a speech recognition module then he has to know if moving from one language to another requires downloading of the appropriate references or if he has to configure and tune various other parameters (thresholds, gains, ...) depending on the second language.

2.11 System dimension configuration

It is important for telephone applications to determine the number of ports required with respect to the "call volume" and the type of information to be delivered. The method of dimensioning system capacity is based on the procedure conducted in the management of call handling resources. It is based on the likely maximum number of calls during the traffic peak period and the average duration of the calls. Of course the number of calls per hour considered here is an average figure and may depend on the activity (opening hours, promotion events, etc.). It may change slightly or significantly with independent external events. Thus, the system is sized to support an average number of calls and to lose a minimum of them. The system may offer a shared resource architecture based on an internally switched access to the required voice processing resource (recognition and/or synthesis) on an as needed basis. In that case one may have fewer recognisers than lines but should apply the following computation to know how many calls will not be satisfied if all resources will be engaged.

- 1. The first parameter is the number of calls/hour. This has to be computed regarding the calls distribution and considering the busiest hour.
- 2. The second parameter is the average call length, which determines the duration of a communication between the application and the caller. It is correlated to the application purpose and its design (message length, cut-through facility, etc.). One may imagine that users carry out a simple information query as well as a complete transaction (a simple one as well as more complex ones). One also has to account for exits that occur usually during the greeting message (people dialling a wrong number or expecting a human operator). So the average call length is computed on the basis of several experiences to minimise inconsistencies and to account for a weighting factor for each type of the scenarios mentioned above (abandon, simple transaction, complex ones).

This computation leads to the complete average call length and the number of calls during the peak period that should be handled by the system. The Erlang is the measure unit that defines the system dimensions. This measure is given by tables showing the number of ports as a function of the percentage of calls that will be lost The table is called the Erlang B

carried-traffic table (see Table 2.5 on page 76).

The unit is defined as follows:

Erlang = $1/3600 \cdot$ Number of calls during the busiest hour \cdot Average call length in seconds

It is obvious that one cannot - especially for economic reasons - provide resources large enough to handle all the calls. It has to be clear that some calls will not be satisfied and will be lost. There is an optimal balance to look for which is a compromise between the number of lost calls and the cost of extra telephone ports. The percentage of lost calls is referred to as the *blockage level*.

Let us assume that during the busiest hour the application should respond to 240 simultaneous calls and that the average call length (the average duration of a call) is 180s.

 $1/3600 \cdot 240 \cdot 180 = 12$ Erlangs.

If one accepts losing one call out of ten (10%), then the closest value in the table to 12 is 11,47, which indicates that the required number of ports is 14.

If one accepts losing only one call out of a thousand (0,1%), then the Erlang table indicates that the closest value is 11,52 and the system should have 23 ports.

Another example: if the number of calls is 480 instead of 240 (two times more) then: to lose less than 10% one needs 26 ports and to loose less than 0.1% one needs 39 ports.

2.12 Statistical tools

The only way to monitor the system consists of call tracking, transaction logfiles, and other traces. The application developer has to know what the available statistics are if he uses an application generator or how he can access them using APIs. He also has to know if there are any tools that allow him to manage such data (daily, weekly, monthly analysis with respect to a given factor).

2.13 Conclusion

This document aims at helping application developers to make sure that the technology (or packaged systems) they choose offer the best capabilities required by their application with respect to some price/performance ratios. For this purpose they should issue a "Requirement profile" to be sent to a large number of vendors. The requirements profile should be structured into two levels: a first level devoted to general information to determine the technologies to be short-listed, and a second level that should allow selection of an appropriate technology partner through an in-depth analysis of the adequacy relation between the application needs and the technologies while saving the time needed for such comparisons.

2.14 Recommendation: Requirement profiles

In this section we will summarise the dimensions of a requirement profile with keywords that will need to be regularly updated and illustrated by potential users.

A. Profile for speech recognition systems

Environment	Noise type (stationary, speech-like?)
	Signal-to-noise ratio
	Utilisation conditions
	Reverberation, acceleration, vibration
Transducer	microphone (open, press-to-talk)
	telephone (mobile?)

System design 71

Channel	bandwidth distortion echo, delay
Task	Lexicon size Lexicon confusability Perplexity factor Dialogue size
Speakers	Speaker dependent/independent number of training speakers available number of training speakers to record typology of users: native/non-native sex, age range physical/psychological state social group attitude (motivated)
Speech	experienced/large public amplitude (quiet/normal/shout) mode (isolated/continuous) fluency (read, spontaneous) rate (slow, normal,fast) conformity (speech/non-speech sounds)
Vocabulary	Training material available Training material to collect? task specific/task independent environment transducer channel users recency control of vocabulary switching
In-situ recording (field data)	possible/not
Rejection mode	

Confusion matrix for the application vocabulary Language modelling Channel and environment adaptation

Task/Application adaptation Speech recognition and application interfaces Speech input and speech signal acquisition Cut-through versus voice-stop (anticipation) Error measure presentation Error handling Response time Analogue connection Analogue to digital conversion Digital interfaces Noise conditions PTT approvals

Controls: Transducer level settings (AGC)

System performance

Recognition error rate real-time aspects system response-time packaging aspects (size, weight, power, cost)

B. Profile for speaker verification systems

Error measure Speaker verification error Training Exploitation Text dependency / independency Speech quantity and quality Adding/Removing speakers C. Profile for a speaker identification system

Error measure Speaker verification error Training Exploitation Text dependency / independency Speech quantity and quality Adding/Removing speakers

D. Profile for a speech synthesis system

Speech recording, storage, and playback Canned speech Text-to-speech synthesis: Linguistic part Phonetic part Acoustic module Multi-linguality

E. EAGLES SLWG keywords for speech synthesis profiles

A set of keywords, defined by the EAGLES Spoken Language Working Group 5 (meeting of November 1–2, 1993, Cambridge, chaired by Louis C.W. Pols) related to system and application characterisations, is given below for information.

1.	Text coverage	From concept, unlimited text, text
		interpretation (e.g. tables),
		(carrier phrases) plus keywords,
		punctuation, spell option, style specification
		language
2.	Source	Coding and/or synthesis and/or canned speech male/female/child
		style, emotion, rate, dialect
		adaptive to disturbances in channel or with user
3.	Channel	High quality, telephone (handset, mobile, ear- phone),
		bandwidth, noise, reverberation, competing speech
4.	User	Experience (one-time vs. multiple use), training, child-normal-elderly, (non-)native, 2nd language
		user,
		hearing impairment, (non-)cooperative

5.	Application	Of reading-machine type OR of information- retrieval type?
		Field test of application itself (task completion)
		laboratory test of synthesis part alone, either with application-specific or application-independent tests.
6.	Functional characteris- tics	Main emphasis on comprehension, intelligibility, naturalness, or otherwise?
		If intelligibility, then of all words or only of certain words?
		Consider separate evaluation of prosodic component
		If overall quality, then use set of scales (see above)
		Consider secondary tasks
		Performance in direct comparison, or in absolute
		sense,
		benchmarking
		How important are dialogue aspects (see other subgroups)?
7.	Restrictions	Time, money, system availability
8.	Alternatives and/or	
	combined modes	Mouse, screen;
		add visual image;
		$\operatorname{multimedia}$?
		Importance of hands-free, eyes-busy?
9.	Technical details	Size, weight, price, interface, plug-in options,
		DSP board, modularity, diphone basis,
		options, hand-tuning, etc.

System design 75

F. System platform	
Software aspects	Operating systems Drivers Application programming interfaces Application generators
Hardware aspects	Platforms Speech processing boards Speech input/output interfaces Connectivity Real-time aspects
Planning for expansion	System simulation and prototyping Host-computer interfaces Computer telephony integration Multi-lingual aspects System dimension configuration Statistical tools

	Blockage factor					
Ports	10%	5%	2%	, 1%	0.5%	0.1%
4	2.05	1.52	1.09	0.87	0.70	0.44
5	2.88	2.22	1.66	1.36	1.13	0.76
6	3.76	2.96	2.28	1.91	1.62	1.15
7	4.67	3.74	2.94	2.50	2.16	1.58
8	5.60	4.54	3.63	3.13	2.73	2.05
9	6.55	5.37	4.34	3.78	3.33	2.56
10	7.51	6.22	5.08	4.46	3.96	3.09
11	8.49	7.08	5.84	5.16	4.61	3.65
12	9.47	7.95	6.62	5.88	5.28	4.23
13	10.47	8.83	7.41	6.61	5.96	4.83
14	11.47	9.73	8.20	7.35	6.66	5.45
15	12.48	10.63	9.01	8.11	7.38	6.08
16	13.50	11.54	9.83	8.87	8.10	6.72
17	14.52	12.46	10.66	9.65	8.83	7.38
18	15.55	13.38	11.49	10.44	9.58	8.05
19	16.58	14.31	12.33	11.23	10.33	8.72
20	17.61	15.25	13.18	12.03	11.09	9.41
21	18.65	16.19	14.04	12.84	11.86	10.11
22	19.69	17.13	14.90	13.65	12.64	10.81
23	20.74	18.08	15.76	14.47	13.42	11.52
24	21.78	19.03	16.63	15.29	14.20	12.24
25	22.83	19.99	17.50	16.12	15.00	12.97
26	23.88	20.94	18.38	16.96	15.80	13.70
27	24.94	21.90	19.26	17.80	16.60	14.44
28	26.00	22.87	20.15	18.64	17.41	15.18
29	27.05	23.83	21.04	19.49	18.22	15.93
30	28.11	24.80	21.93	20.34	19.04	16.68
31	29.17	25.77	22.83	21.19	19.86	17.44
32	30.23	26.75	23.73	22.05	20.68	18.20
33	31.30	27.72	24.63	22.91	21.51	18.97
34	32.36	28.70	25.53	23.77	22.34	19.74
35	33.43	29.68	26.43	24.64	23.17	20.52
36	34.50	30.66	27.34	25.51	24.01	21.30
37	35.57	31.64	28.25	26.38	24.85	22.08
38	36.64	32.63	29.17	27.25	25.69	22.86
39	37.71	33.61	30.08	28.13	26.54	23.65
40	38.79	34.60	31.00	29.01	27.38	24.44
41	39.86	35.59	31.92	28.89	28.23	25.24
42	40.94	36.58	32.84	30.77	29.08	26.04

Table 2.5: Erlang B carried-traffic table

System design 77

			Blockag	e factor		
Ports	10%	5%	2%	, 1%	0.5%	0.1%
43	42.01	37.57	33.76	31.66	29.94	26.84
44	43.09	38.56	34.68	32.54	30.80	27.64
45	44.16	39.55	35.61	33.43	31.66	28.45
46	45.24	40.54	36.53	34.32	32.52	29.26
47	46.32	41.54	37.46	35.21	33.38	30.07
48	47.40	42.54	38.39	36.11	34.25	30.88
49	48.48	43.54	39.32	37.00	35.11	31.69
50	49.56	44.53	40.25	37.90	35.98	32.51
51	50.60	45.50	41.20	38.80	36.85	33.30
52	51.70	46.50	42.10	39.70	37.72	34.20
53	52.80	47.50	43.10	40.60	38.60	35.00
54	53.90	48.50	44.00	41.50	39.47	35.80
55	55.00	49.50	44.90	42.40	40.35	36.60
56	56.10	50.50	45.90	43.30	41.23	37.50
57	57.10	51.50	46.80	44.20	42.11	38.30
58	58.20	52.60	47.80	45.10	42.99	39.10
59	59.30	53.60	48.70	46.00	43.88	40.00
60	60.40	54.60	49.60	46.90	44.76	40.80
61	61.50	55.60	50.60	47.90	45.64	41.60
62	62.60	56.60	51.50	48.80	46.53	42.50
63	63.70	57.60	52.50	49.70	47.42	43.30
64	64.80	58.60	53.40	50.60	48.31	44.20
65	65.80	59.60	54.40	51.50	49.19	45.00
66	66.90	60.60	55.30	52.40	50.09	45.80
67	68.00	61.60	56.30	53.40	50.98	46.70
68	69.10	62.60	57.20	54.30	51.87	47.50
69	70.20	63.70	58.20	55.20	52.77	48.40
70	71.30	64.70	59.10	56.10	53.66	49.20
71	72.40	65.70	60.10	57.00	54.56	50.10
72	73.50	66.70	61.00	58.00	55.46	50.90
73	74.60	67.70	62.00	58.90	56.35	51.80
74	75.60	68.70	62.90	59.80	57.25	52.70
75	76.70	69.70	63.90	60.70	58.15	53.50
76	77.80	70.80	64.90	61.70	59.05	54.40
77	78.90	71.80	65.80	62.60	59.96	55.20
78	80.00	72.80	66.80	63.50	60.86	56.10
79	81.10	73.80	67.70	64.40	61.77	57.00
80	82.20	74.80	68.70	65.40	62.67	57.80
81	83.30	75.80	69.60	66.30	63.57	58.70

Table 2.6: Erlang B carried-traffic table (continued)

			Blockage	factor		
Ports	10%	5%	2%	1%	0.5%	0.1%
82	84.40	76.90	70.60	67.20	64.48	59.50
83	85.50	77.90	71.60	68.20	65.39	60.40
84	86.60	78.90	72.50	69.10	66.29	61.30
85	87.70	79.90	73.50	70.00	67.20	62.10
86	88.80	80.90	74.50	70.90	68.11	63.00
87	89.90	82.00	75.40	71.90	69.02	63.90
88	91.00	83.00	76.40	72.80	69.93	64.70
89	92.10	84.00	77.30	73.70	70.85	65.60
90	93.10	85.00	78.30	74.70	71.75	66.50
91	94.20	86.00	79.30	75.60	72.67	67.40
92	95.30	87.10	80.20	76.60	73.58	68.20
93	96.40	88.10	81.20	77.50	74.49	69.10
94	97.50	89.10	82.20	78.40	75.41	70.00
95	98.60	90.10	83.10	79.40	76.33	70.90
96	99.70	91.10	84.10	80.30	77.24	71.70
97	100.80	92.20	85.10	81.20	78.16	72.60
98	101.90	93.20	86.00	82.20	79.07	73.50
99	103.00	94.20	87.00	83.10	79.99	74.40
100	104.10	95.20	88.00	84.10	80.91	75.20
101	105.20	96.30	88.90	85.00	81.83	76.10
102	106.30	97.30	89.90	85.90	82.75	77.00
103	107.40	98.30	90.90	86.90	83.67	77.90
104	108.50	99.30	91.90	87.80	84.59	78.80
105	109.60	100.40	92.80	88.80	85.51	79.70
106	110.70	101.40	93.80	89.70	86.43	80.50
107	111.80	102.40	94.80	90.70	87.36	81.40
108	112.90	103.40	95.70	91.60	88.28	82.30
109	114.00	104.50	96.70	92.50	89.20	83.20
110	115.10	105.50	97.70	93.50	90.12	84.10
111	116.20	106.50	98.70	94.40	91.05	85.00
112	117.30	107.50	99.60	95.40	91.97	85.90
113	118.40	108.60	100.60	96.30	92.89	86.70
114	119.50	109.60	101.60	97.30	93.82	87.60
115	120.60	110.60	102.50	98.20	94.74	88.50
116	121.70	111.70	103.50	99.20	95.67	89.40
117	122.80	112.70	104.50	100.10	96.60	90.30
118	123.90	113.70	105.50	101.10	97.52	91.20
119	125.00	114.70	106.40	102.00	98.45	92.10
120	126.10	115.80	107.40	103.00	99.38	93.00

Table 2.7: Erlang B carried-traffic table (continued)

3 SL corpus design

3.1 Introduction

3.1.1 Spoken language corpus

The definition introduced here for a spoken language corpus is "any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data in-house, or by scientists in other organisations." This tentative definition excludes a large number of speech recordings on analogue tapes (sometimes even on disks) and recordings without the annotation and documentation which is necessary in order to use the recordings effectively. For instance, it is well known that virtually all public broadcasting corporations in Europe maintain an archive of recordings of programmes, including newscasts, reports of events ranging from football matches to royal weddings and funerals. However, in most cases these recordings can only be accessed by the date of the original broadcast, and perhaps also by the type of programme. Only in very rare cases are transcripts of the speech material in the recordings available. This makes it extremely difficult and time-consuming to use these data for almost all types of research. Speech coding forms the most notable exception to this rule, although even for coding research knowledge of who has said what may be helpful. Of course, lack of annotation does not diminish the value of these recordings for cultural and scientific purposes, but due to the inordinate amount of pre-processing necessary for any type of research they do not qualify as a spoken language corpus under our definition. In many other respects our definition is very wide and liberal. For instance, a set of computer files containing speech signals, EMG signals and sub- and supraglottal pressure signals measured in two subjects who sustained vowels on different pitch and intensity levels would qualify as a *spoken language corpus*, provided that the files come with suitable annotation and documentation.

Many additional sources of information can be gathered apart from the basic acoustic speech signal. Whatever choices of speakers, speech material, and recording conditions are made, it is always of crucial importance that the collecting procedure is documented as elaborately as possible. It is good practice to record all possible details about, for instance, sex (gender) and age of speakers, type of speech material (isolated words, sentences, discourse, etc.), place of recording (in a laboratory, on location, etc.), type of microphone and recording medium (see also Chapter 8). Although one may not be interested in specific information at the time, it can turn out to be important at a later stage. In that case it is often difficult or impossible to recollect the required information. Furthermore, a well documented speech
corpus may also be used for other directions of research. The following list summarises the most common information sources that may be present in a speech corpus:

- Transduced signals
- Examples: The acoustic speech signal, laryngograph signal, X-ray data. • Analysis results
- Examples: FFT data, LPC data, filter bank data, pitch extraction, formant extraction.
- Descriptors
- Examples: Characteristics of the speakers, or the recording conditions. • Markers
- Examples: Markers to indicate pitch periods, or the beginning of vowels. • Annotations/Labels
- Examples: Orthographic, phonemic, or phonetic transcriptions.
- Assessment parameters Examples: Test material, assessment results.

All these information sources must be stored in such a way that potential users of the speech corpus can get access to the speech and the speechrelated data in an efficient and easy-to-use manner.

3.1.2 About this chapter

In this chapter only the *design* of spoken language corpora and the *use* of these corpora are covered. It is *expressis verbis* not the intention of this chapter to give a comprehensive overview of corpora existing worldwide; we even do not intend to give a comprehensive list of all corpora existing in Europe. For attempts to survey existing speech corpora the reader is referred to Fourcin et al. (1989) and to Appendix L which contains a list of existing public domain spoken language corpora.

Corpora, tools, and resources in general are not aims in their own right, but means to an independently specified purpose. Thus, the eventual specification of a corpus depends in an essential way on the purpose it is intended to serve. Yet, if that purpose is not too limited, and provided the corpus is properly documented and annotated, it is quite likely that it will be useful for other, perhaps unrelated research. At present there are few, if any, official standards for corpus development. Given the dependence on research goals, this is not surprising.

The present chapter intends to address as large an audience as possible. Specifically, it includes information and recommendations not only for speech technology research, but also for the development of corpora meant to support research in speech science, psycholinguistics and sociolinguistics. The recommendations concern general aspects and factors that should be considered in designing a corpus, and guidelines for making decisions on these issues.

In the *pre-recording phase* one has to define the content of the corpus. Specifications of experiment design, of linguistic content, of number and type of speakers, and of the physical situation must be established. These topics will be covered in this chapter. In the *recording phase* speaker instruction and prompting, experiment and recording control, as well as storage of the recordings are involved. These topics will be covered in Chapter 4. In the *post-recording phase* transcription (and possibly segmentation and labelling), corpus lexicon construction, and database management take place. These topics will be discussed in Chapters 5 and 6.

In the remainder of this chapter we focus on the *pre-recording phase*, including the following steps in preparing the recording of a speech corpus:

- defining the application of the corpus,
- specifying the linguistic content of the corpus,
- specifying the number and type of speakers.

Before we embark on these discussions, however, it is necessary to elaborate on the differences between *written language corpora* and *spoken language corpora*.

3.2 Eight main differences between collections of written and spoken language data

Traditionally, linguists and natural language processing (NLP) researchers understood language corpora to consist of written material collected from text sources which already exist and often are available in published form (novels, stage and screen plays, newspapers, manuals, etc.). In this context the term "spoken language text corpora" was used to indicate that the data are not taken from existing texts but that speech had to be written down in some orthographic or non-orthographic form in order to become part of a data collection. However, the differences (and relations) between text and speech data are far more complex. There are at least eight important differences, which must not be ignored because they determine relevant properties of the resulting data collections. For future (technological) developments of Spoken Language Processing (SLP) they should be taken into account very seriously.

These eight differences have to do with:

- 1. the durability of text as opposed to the volatility of speech,
- 2. the different time it takes to produce text and speech,
- 3. the different roles errors play in written and spoken language,
- 4. the differences in written and spoken words,
- 5. the different data structures of ASCII strings and sampled speech signals,
- 6. the two reasons that cause the great difference in the size of NL and SL data collections,

- 82 Part I: Spoken language system and corpus design
- 7. the different legal status of written text and spoken utterances, and
- 8. the most fundamental distinction (as well as relation) between symbolically specified categories and physically measured time functions.

A closer look at these eight differences between written and spoken data will reveal why the traditional term "natural language processing", NLP, also could well be read as standing for "Non-spoken Language Processing". As it is our goal to call special attention to the relevant differences we will refer to the written language data as NL data meaning *non-spoken language data*, and set it in opposition to the term SL data, the acronym for *spoken language data*.

3.2.1 Durability of text, volatility of speech

The first distinction may seem rather trivial but it must nonetheless be mentioned, because it affects specific properties of the collected NL and SL data. While text generally stays on the paper when it is written down, speech is transient. It is the nature of the phonetic facts which speakers create during speech acts that they disappear at the moment they come into existence.

The first difference (which in the step from speaking to writing has helped our cultural development) explains why to collect SL data is less trivial than to produce NL data. The former must necessarily be recorded, for example on a tape or a disk, to make it accessible for future use.

3.2.2 Different production times for text and speech

Another difference between NL and SL corpora is due to the fact that speech data are time functions in a sense in which text data are not. Whilst a writer may consume any time he wants (or needs) to invest in producing a text, a speaker must code and transmit the phonetic information through syllabically and rhythmically organised sound transitions. Speech must run in its own natural time with a typical syllable rate of a value between 120/min and 180/min. The time for writing new text is normally much longer than it takes to read it aloud (which does not mean that silent reading and shorthand-writing cannot be much faster than speaking the text).

3.2.3 Correcting errors in the production of text and speech

In spontaneously spoken language the editing behaviour of the speaker is audible and remains a part of the recorded data. Interruptions, hesitations, repetitions of words (and parts of words), and especially self-repairs are a characteristic feature of naturally spoken language and must be represented in SL data collections of spontaneous speech. On the other hand, the writer who has even more correcting and editing options in producing a text document, will normally intend to produce a "clean" version. In the final version of the text all corrections which may have been carried out have disappeared; this is especially true for text intended to go into print. In the recent past SL data were often recorded as clean speech collections. A typical example is so-called laboratory speech which is produced when a speaker who is sitting in a monitored recording room reads a list of prepared text material, and then only the proper reproductions of the individual text items are accepted to enter the data base. Examples of speech corpora collected in this way are EUROM-0 and EUROM-1 (see Appendix J). More recently, however, interest has shifted towards corpora comprising "real-world" speech, including hesitations, corrections, background noise, etc.

3.2.4 Orthographic identity and phonetic variability of lexicalised units

In correctly written texts any morphologically inflected lexical item generally has just one distinct orthographic form. Thus the words of European languages are easily identified and also well distinguished from each other, and there is usually only one version of each possible orthographic contextual form of any given word. The spoken versions of orthographically identical word forms show a great phonetic variation in their segmental and prosodic realisation. In most European languages the phonetic form of a given word is in fact extremely variable depending on the context and other well defined intervening variables such as speaking style and context of situation, strong and weak Lombard effects (the influence of the physical environment on speech production via acoustic feedback), etc. A given word can totally disappear phonetically, or can be reduced to – and only signalled by – some reflection of segmental features in the prosody of the utterance. Most of these inconspicuous variations appear only in a narrow phonetic transcription of a given pronunciation.

It makes a great difference whether a word has been uttered in isolation or in continuous speech. Only if a word is consciously and very carefully produced in isolation can we observe the explicit version of its segmental structure. These phonetically explicit forms produced in a careful speaking style are called citation forms or canonical forms. The segmental structure of so-called citation forms is modified as soon as it is integrated into connected speech (probably systematically, although relatively little of the system is currently understood). For the design of spoken language corpora this is very relevant. It has also been taken into account in the conventions of the IPA proposed for Computer Representation of Individual Languages (CRIL, see Appendix A).

In dealing with SL data one must be able to know which words the speaker intended to express in a given utterance. This is reflected in the CRIL convention of the IPA (see Section 5.2.4). Here it should be mentioned that an SL data collection should ideally have at least two and possibly three dif-

ferent symbolically specified levels which are related to the acoustic speech signal:

- 1. On the first level the words of the given utterance are identified as lexical units in their orthographic form.
- 2. On the second level a broad phonetic transcription of the citation form should be given, which may be the result of automatic grapheme-tophoneme conversion, as for very large SL corpora it would cost too much time and too much money to make broad phonetic transcriptions manually.
- 3. How the given words have been actually pronounced in a given speech signal must be specified in terms of a narrower phonetic transcription of each individual utterance on a third, optional CRIL-level. This third level can then be directly aligned to the segments or acoustic features of the digital speech signal in the data base, which can be done automatically or manually. This information is especially relevant if also multi-sensor data are to be incorporated in SL databases.

Detailed phonetic transcriptions are subject to intra and inter-transcriber variability. Furthermore, they are extremely expensive, to the extent that they are likely to be prohibitive for large corpora. However, recent attempts using large vocabulary speech recognisers for the acoustic decoding of speech show some promise that the process can be automated, at least to the extent that pronunciation variation can be predicted by means of general phonological and phonetic rules.

In addition to phonetic detail on the segmental level, several uses of spoken language corpora may also require prosodic annotation. In this area much work remains to be done to develop commonly agreed annotation systems. Once such systems exist, one may attempt to support annotation by means of automatic recognition procedures.

3.2.5 Printable ASCII-strings and continuously sampled speech

Taken as pure data, written texts in European languages consist of strings of printable alphanumerical and other elements coded in 7- or 8-bit ASCII-Bytes. The resulting NL strings possess already a characteristic information structure which is not available in the case of primary SL data. Separated by blanks, punctuation marks or control codes, ASCII-strings are grouped into lexical substrings; also, the explicit punctuation of phrases and sentences is an important property of NL data. None of this type of information can be found in the recordings of primary SL data, since in natural speech there are no ASCII elements representing word boundaries, full stops, commas, colons, quotation, question, exclamation marks. Recorded SL data are primarily nothing but digitalised time functions.

3.2.6 Size differences between NL and SL data

Comparing the pure size of stored NL and SL data reveals a great quantitative difference. There are two reasons why SL data require orders of magnitude more storage space than written language corpora. The first one is simply the difference in coding between text and speech. Whereas the ASCII string of a word like *and* needs only three bytes, many more bytes are required as soon as the phonemes of this word are transformed into an acoustic output for storing the AD-converted data. If in the given example we assume that in clear speech the utterance of a three-phoneme-syllable takes about half a second and if we apply an amplitude quantisation of 16 bits and a non-stereo hi-fi sampling rate of 48 kHz, the NL/SL ratio amounts to approximately 1:16000.

The second reason follows from the great variability in the phonetic forms of spoken words. As pointed out above, any written text must be reproduced by many speakers in more than one speaking style (at least at slow, normal and fast speeds with low, normal, high voice, etc.), if the corpus is intended to reflect some common sources of variability.

3.2.7 The different legal status of written texts and spoken words

With few exceptions, the texts in NL corpora have previously been published. From a legal point of view, this implies that any use of electronic copies should adhere to copyright rules and regulations. In most countries copyright laws were passed long before the era of electronic publishing. However, laws designed to protect printed materials may not be optimal for the protection of machine readable text. Neither is it obvious how abuse of electronic texts can be detected and prevented. These problems have impeded the distribution of NL corpora quite considerably and it would be optimistic to suggest that all problems are close to a solution.

For SL corpora the legal issues are even less well understood. Has a speaker who is recorded while reading sentences presented by an experimenter any legal rights with respect to the sounds produced? Recordings of spontaneous speech are even more complex in this respect, since a speaker might claim rights as to the contexts and details of the formulations used. If speakers are recruited to contribute to a SL corpus, legal problems can be avoided by requesting them to sign a consent form. Building corpora from existing recordings (e.g. from radio and television broadcasts) is more difficult in this respect, because it may not always be feasible to contact all relevant speakers. Under the law of EU countries unauthorised re-broadcast of recordings made from radio or television is illegal. It is less clear what the legal status is of limited redistribution of recordings for research and development in speech science and technology. For more information on this topic, we refer to Section 4.3.4.

3.2.8 The different nature of categories and time functions

The last difference, and the most important one, must be looked at from two different angles. The first thing to understand is that the relevant category of the data (that determines its collection) is already inherently given in the case of NL, but totally unknown in the case of physically recorded speech. The ASCII symbols of a given text are elementary categories by themselves, and are directly used to form syntactically analysable expressions for the representation of all the different linguistically relevant categories. Thus relevant categorical information can be directly inferred from categorically given data and their ASCII representations. In contrast to this NL situation, the data of a digital speech signal do not signal any such categories, because they only represent a measured time function without any inherent categorical interpretation. At the present stage in the development of SLP it is not yet even possible to decide automatically whether a given digital signal is a speech signal or not. Therefore the necessary categorical annotations for SL data must still be produced by human workers (with the increasing support of semi-automatic procedures).

The second matter that must be considered in judging the different roles of categories and time functions in speech technology is that speech signals contain relevant prosodic and paralinguistic information that is not represented by the pure text of what was pronounced within a given utterance. As long as NLP can be restricted to non-spoken language processing the restriction to NL data does not pose severe problems. But as soon as real speech utterances are to be processed in an information technology application, the other, non-linguistic, but communicatively extremely relevant categories cannot be ignored. They must be represented in future SL data collections, and much effort has still to be invested by the international scientific community to deal with all these information-bearing aspects of any given speech utterance.

3.3 Applications of spoken language corpora

As mentioned earlier, speech corpora are always designed for specific purposes. These purposes determine the content and design of a corpus. Thus, a speech therapist interested in pathological speech will collect a completely different corpus than a designer of a telephone response application. For example, in the first case hi-fi speech recordings are most probably needed in order to study properties of voice quality, whereas in the latter case realistic speech should be collected over the phone, which will result in a rather poor quality of the speech.

In this section we will present a non-exhaustive list of possible users of speech corpora together with the specific types of speech corpora they would need. A distinction will be made between corpora for research purposes and those

meant for technological applications. Of course, this does not mean that corpora gathered in the one field cannot be used in the other, although there will be differences in the exchangeability of corpora depending on the corpus. It must be clear that we cannot handle all the details of specific corpora, and that we will indicate only some general properties.

3.3.1 Speech corpora for research purposes

The speech corpora needed for scientific purposes can be very diverse. Some researchers may need carefully pronounced lists of words to study a specific hypothesis about speech production; others may want to study samples of the vernacular, the way people speak in their everyday life. In the following sections some of the major scientific fields with interest in spoken language corpora are mentioned.

3.3.1.1 Phonetic research

In phonetic research all aspects of speech are studied. Phonetic experiments often require carefully controlled speech data, especially when basic phenomena, such as coarticulation, have to be studied in a systematic way. In this type of research, more often than not the researcher will have no alternative to collecting new data, specifically designed for the investigation at hand. However, in recent years more and more attention is being paid to uncontrolled (or less controlled) forms of speech as well, because one has begun to realise that results obtained for carefully pronounced speech cannot simply be generalised to more casual speech. This type of research, which requires other experimental designs and other statistical test procedures as well, will profit considerably from existing corpora. Moreover, since the corpora that can support this type of research must of necessity be very large, it will be very unlikely that a researcher will have the opportunity to collect new, project-specific corpora.

3.3.1.2 Sociolinguistic research

In sociolinguistic research variation in language use is studied in heterogeneous communities, especially urban ones. Variables of interest are among others age, sex (gender), and social status. Three common methods to gather data in this research field are:

• By means of written questionnaires

Members of the communities of interest might, for instance, have to indicate on a questionnaire how they pronounce specific words, or whether they use certain sociolect variants of words. A large drawback of this method is that many people are not aware of their pronunciation habits, or the sociolect variants they use. Furthermore, people might regard their actual language use as undesirable and vulgar, and pretend they use a more prestigious form of language.

• By observations of the investigator

This strategy was, for instance, used by William Labov to investigate the occurrence of /r/ deletions in New York English (Labov 1972). He simply wrote down whether his informants pronounced an /r/ or not in specific words. The major drawback of this method is that the data collection is based on the subjective (and possibly biased) observations of a single person. In addition, the phenomenon of interest is only heard once at a possibly unexpected moment and in a possibly noisy environment (Labov, for instance, did an investigation in department stores).

• By collecting speech corpora

The gathering of speech corpora offers sociolinguists the opportunity to make detailed and reliable analyses of various phenomena of interest. Perceptual evaluation of pronunciation phenomena could be supplemented or replaced with acoustic measurements. This is especially useful when the differences between pronunciation variants are very subtle, as for instance in the case of a slightly varying vowel colour (Labov 1994).

Dialect research is closely related to sociolect research. In dialect studies variation in language use due to differences in geographical background of speakers is investigated. Since the methods of data collection are similar to the ones used in sociolect research, the remarks made above also apply to dialect research.

3.3.1.3 Psycholinguistic research

Psycholinguistics is a very broad scientific field in which the psychology of language is studied, from language acquisition by children to the mental processes underlying adult comprehension and production of speech and language disorders.

Psycholinguistic experiments sometimes involve carefully controlled speech material, for instance in on-line phoneme monitoring or gating experiments. In phoneme monitoring experiments subjects are asked to spot the first occurrence of a specific phoneme in a spoken utterance, and press a button as soon as they have spotted it. The reaction time between the actual occurrence of the phoneme and the subject's response is used to form hypotheses about underlying mental processes. In gating experiments a progressively larger portion of words is presented to listeners, who are asked to predict what the ending will be. Both techniques can be useful to get more insight into the organisation of the mental lexicon (Aitchison 1994).

Another way to obtain information about the mental lexicon and speech production processes is to study the dysfluencies in spontaneous speech. For example, false starts tell us something about the way in which speech is planned and articulated. Also repetitions of words or word fragments give information about the production and representation of speech. For this type of research, spontaneous speech corpora are very useful. For more information about planning processes of speech see Levelt (1989). Yet another way to gather cues about the mental lexicon is to study "slips of the tongue". Many tongue-slip collectors carry round a small notebook in which they write down errors whenever they hear them, on a bus, at parties, etc. As mentioned in the former section, data acquired in this way is subjective and unreliable. The use of speech corpora containing spontaneous speech samples would be the answer to this problem, but investigations in this research area would only benefit from extremely large spontaneous speech corpora, because the number of slips of the tongue produced in any one hour of spoken speech is fairly small.

3.3.1.4 First language acquisition

Language acquisition by children is subject of investigation in many disciplines of, among others, linguistics and psychology. For example, the speech of (young) children can be used to investigate (ir)regularities in language (*linguistics*); it can also be used to learn more about the mental organisation of language (*psycholinguistics*); it can be studied in relation to the *sociolinguistic* background of children; or it can be used to gain more insight into basic *phonetic* processes. All these scientific fields as well as early learning oriented technologies would benefit from extensive corpora containing speech of children.

Collecting language acquisition corpora is extremely time consuming and expensive, because of the difficulty in transcribing the speech, especially speech of very young children. In (psycho-)linguistics a considerable amount of work has been done to collect and transcribe corpora, and to make them available to the research community. Presently, only transcriptions are readily accessible (e.g. the CHILDES transcription of MacWhinney 1995).

In the case of toddlers only "spontaneous" speech samples can be obtained. As soon as children get somewhat older, more controlled forms of speech can also be obtained, such as naming pictures or reading texts. Game playing is another way of eliciting quasi-controlled speech.

Speech acquisition corpora must preferably be longitudinal, i.e. the same person must be recorded repeatedly at consecutive stages in the acquisition process.

3.3.1.5 Second language acquisition

Migration between language areas is as old as history, and no doubt much older. Depending on their practical and social status, migrants may be hindered by their lack of adequate knowledge and command of the majority language in their new home countries. Now that low-education jobs are becoming increasingly rare in First World countries this situation has become economically and politically significant. Because command of the language is a prerequisite to education, the study of how immigrants learn to master the language of the host country (the "second" language) has become an

important topic in sociolinguistic research. The European Science Foundation, for instance, has sponsored a large scale project on second language learning in several Western European countries. The research was corpus based: large numbers of migrants were recorded each fortnight for over a year. Transcripts and audio tapes comprising this corpus are maintained by the Max Planck Institute of Psycholinguistics in Nijmegen, The Netherlands.

It is especially important to study second language acquisition of immigrant children in order to find out how this might influence their education progress. In a similar vein, research into the acquisition of the majority language is needed in "second generation children" who grow up in families which still use the language of their country of origin.

Since immigrants form a minority group in the country they reside in, their native language can be strongly influenced by the second language. For the investigation of these so-called language attrition processes special purpose corpora must be (and have been) collected. In this context one must not only think of African and Asian migrants who are living in the U.S.A. or Western Europe, but also of non-Anglo Europeans who moved to the U.S.A., Canada or Australia.

From a psycholinguistic point of view, it is interesting to study how the different lexicons are organised in the minds of bilingual (and multilingual) speakers. For example, the occurrence of "blends" (combinations of two words, in this case from different languages) shows that words are subconsciously activated in both languages (Green 1986). Up to now, much of the research into bilingual lexicons has taken the form of controlled experiments (e.g. cross language priming in lexical decision tasks). It is conceivable, though, that large corpora of spontaneous speech of bilinguals could be used to study lexical and syntactic interferences between the languages. Large corpora of speech of second language learners may also be very interesting for the development of tools for second language learning. For example, types of grammatical and pronunciation errors can be identified. Knowledge of these errors may be helpful for the development of language instruction materials, which might include spoken examples of actual errors (to be corrected by the learners).

3.3.1.6 General linguistic research

A substantial part of modern linguistic research since the 1960s has been based on Chomsky's "generative paradigm". The goal of this so-called *mentalistic* research programme is to eventually understand the *competence* of language users, i.e. their abstract knowledge of the language system. What speakers and hearers actually do, i.e. their *performance*, is usually of less interest to linguists in Chomsky's tradition. The construction of competence models is generally based on introspection and impressionistic ideas about language use. So, in its strictest form mentalistic linguistic research cannot benefit much from speech corpora that contain samples of the performance of language users.

However, many linguists no longer think that performance can be neglected completely. For one thing, it has been noted that spontaneous speech corpora often contain utterances which would seem implausible (if not impossible) from introspection, but which are perfectly natural and acceptable in context. And conversely, sentences invented to illustrate grammatical points may be implausible as actual utterances, because it is extremely difficult to imagine a situation in which they would not violate discourse constraints, aspectual perspectives taken on events, etc. (Chafe 1992). Moreover, only an *integrated* theory of competence and performance would ultimately be able to account for actual language phenomena. In this respect speech corpora are indispensable to fill the gap between a competence grammar and actual language use.

Presently, more and more linguists are starting to realise the importance of linguistic analysis of constructs of larger size than isolated sentences or utterances. *Discourse analysis* is the branch of linguistics which is concerned with the analysis of naturally occurring connected spoken or written discourse (Stubbs 1984). Obviously, discourse analysis will profit very much from large corpora of meaningful speech, whether it is conversational or more formal, e.g. in information seeking dialogues.

In Edwards and Lampert (1993) a comprehensive methodology is presented for the transcription and coding of discourse data from various perspectives. This book also contains a list of language corpora that might be useful in discourse research.

3.3.1.7 Audiology

Audiology is the scientific study of *hearing*, often including the treatment of persons with hearing defects. A conventional *audiometer* can be used to test the intensity and frequency range of pure tones that the human ear can detect. This instrument can give a rough indication of the degree of hearing loss in hearing-impaired persons. Present day evaluation of hearing includes the use of controlled speech samples to assist in the determination of a patient's communicative capabilities.

Interest in the use of speech to measure hearing has been centered around both research orientation and practical clinical orientation. The first orientation has resulted in research areas such as experimental phonetics, the effects of various types of distortion on human speech recognition and speaker identification, etc. The second orientation has led to research in areas such as the effects of hearing loss on the reception of speech, auditory processing, and the effects of modifications in the range of reception of speech. The

second area more or less grew out of the research in the first area (O'Neill 1975).

For speech corpora to be useful in audiology they must be carefully calibrated, establishing performance (e.g. in terms of recognition scores) of non-hearing impaired reference subjects. Audiological test corpora may contain various types of speech stimuli to evaluate normal and disordered hearing acuity. The speech stimuli can consist of isolated phonemes, nonsense words or real words, and also of connected forms of speech (see e.g. House et al. 1965; Voiers 1977).

3.3.1.8 Speech pathology

In this scientific field various types of pathological speech are studied, ranging from mild disorders such as hoarseness to severe disorders such as aphasia. The aim of most studies of pathological speech is to find therapies that can alleviate or cure the speech disorder of interest. However, phenomena like aphasia can also be subject of psycholinguistic studies, because such language disorders can shed some light on underlying mental processes (Aitchison 1994). Corpora of pathological speech are very useful for these purposes. These corpora may also be useful for the development of automatic classifications of speech pathologies.

3.3.2 Speech corpora for technological applications

Technological applications for which speech corpora are needed can be roughly divided into four major classes: speech synthesis, speech recognition, spoken language systems, and speaker recognition/verification. Depending on the specific application, the speech corpora which are needed are very diverse. For example, speech synthesis usually requires a large amount of speech data from one or two speakers, whereas speech recognition often requires a smaller amount of speech data from many speakers. In the following sections the four domains of speech research for technological applications and the speech corpora they need are discussed.

3.3.2.1 Speech synthesis

The seemingly most natural way to synthesise speech is to model human speech production directly by simulating lung pressure, vocal fold vibration, articulatory gestures, etc. However, the human system is not completely understood. This is probably the reason why it turns out to be extremely difficult to determine and control the details of the model parameters in computer simulations. This is the reason that articulatory synthesisers have only been moderately successful in generating perceptually important acoustic features. Yet, modern measurement techniques have allowed the collection of substantial amounts of measurement data. Most of these data are now being made available to the research community (see the ESPRIT project ACCOR (ESPRIT/BRA 3279 ACCOR, and the special issue of the journal "Language and Speech" (1993)).

A relatively simple way to build a speech synthesiser is through concatenation of stored human speech components. In order to achieve natural coarticulation in the synthesised speech, it is necessary to include transition regions in the building blocks. Often-used transition units are diphones, which represent the transition from one phone to another. Since diphone inventories are derived directly from human utterances, diphone synthesis might be expected to be inherently natural sounding. However, this is not completely true, because the diphones have to be concatenated and in practice there will be many diphone junctions that do not fit properly together. In order to be able to smoothe these discontinuities the waveform segments have to be converted to a convenient format, such as some form of LPC parameters, often with some inherent loss of auditory quality. Until recently it was believed that a parametric representation was mandatory to be able to change the pitch and timing of utterances without disturbing the spectral envelope pattern. Since the invention of PSOLA-like techniques, high quality pitch and time changes can be effected directly in the time domain. For limited applications, such as train information systems, whole words and even phrases may be stored. Lately, this method of speech synthesis is being applied more and more, because of the possibility of cheap mass storage. The quality of concatenated-word sentences is often acceptable, especially in the light of the still not optimal quality of the other types of synthesis.

Another important method of generating computerised speech is through *synthesis by rule*. The usual approach is to input a string of allophones to some form of formant synthesiser. Target formant values for each allophone are derived from human utterances and these values are stored in large tables. With an additional set of rules these target values can be adapted to account for all kinds of phonological and phonetic phenomena and to generate proper prosody.

More detailed accounts of speech synthesis systems can be found in, for instance, Klatt (1987) and Holmes (1988), and in Chapter 12.

For all types of speech synthesis systems corpora are needed to determine the model parameters. If the user wants many different types of voice, the speech corpus should contain various speakers for the extraction of speakerspecific model parameters. In particular, the user might want to be able to generate both male and female speech. Transformations to convert rule systems between male and female speech have had limited success, so it seems more convenient to include both sexes in the speech corpus. Application specific corpora are needed to investigate issues related to prosody.

3.3.2.2 Speech recognition

There are several types of speech recognition systems, which may differ in three important ways:

- 1. the recognition strategies they use,
- 2. the speakers they have to recognise,
- 3. the speech they have to recognise.

These different aspects will be discussed below.

3.3.2.2.1 Knowledge-based vs. stochastic systems

With respect to the strategies they use, speech recognition systems can be roughly divided in two classes: *knowledge-based* systems and *stochastic* systems. All state-of-the-art systems belong to the second category. In the knowledge-based approach an attempt was made to specify explicit acousticphonetic rules that are robust enough to allow recognition of linguistically meaningful units and that ignore irrelevant variation in these units. Stochastic systems, such as Hidden Markov Models (HMMs) or neural networks, do not use explicit rules for speech recognition. On the contrary, they rely on stochastic models which are estimated or *trained* with (very) large amounts of speech, using some statistical optimalisation procedure (e.g. the Estimate-Maximise or the Baum-Welch algorithm).

Higher level linguistic knowledge can be used to constrain the recognition hypotheses generated at the acoustic-phonetic level. Higher level knowledge can be represented by knowledge-based explicit rules, for example syntactic constraints on word order. More often it is represented by stochastic language models, for example *bigrams* or *trigrams* that reflect the likelihood of a sequence of two or three words, respectively (see also Chapter 7). Recently, promising work on enhancing HMMs with morphological and phonological structure has been conducted, pointing to the possibility of convergence between knowledge-based and stochastic approaches.

3.3.2.2.2 Speaker-independent vs. speaker-dependent systems

Speech recognition systems can be either *speaker-dependent* or *speaker-independent*. In the former case the recognition system is designed to recognise the speech of just a single person, and in the latter case the recognition system should be able to recognise the speech of a variety of speakers. All other things being equal, the performance of speaker-independent systems is likely to be worse than in speaker-dependent systems, because speaker-independent systems have to deal with a considerable amount of inter-speaker variability. It is often sensible to train separate recognition models for specific subgroups of speakers, such as men and women, or speakers with different dialects (Van Compernolle et al. 1991).

Some systems can to some extent adapt to new speakers by adjusting the parameters of their models. This can be done in a separate training session with a set of predetermined utterances of the new speaker, or it can be done on-line as the recognition of the new speaker's utterances gradually proceeds.

Most recognition systems are very sensitive to the recording environment. In the past, speakers employed to train and develop a system were often recorded under "laboratory" conditions, for instance in an anechoic room. It appears that the performance of speech recognisers which are trained with such high quality recordings severely degrades if they are tested with some form of "noisy" speech (Gong 1995). Also the use of different microphones during training sessions and test sessions has a considerable impact on recognition performance.

3.3.2.2.3 Isolated words vs. continuous speech

The third main distinction between speech recognition systems is based on the type of speech they have to recognise. The system can be designed for *isolated word recognition* or for *continuous speech recognition*. In the latter case word boundaries have to be established, which can be extremely difficult. Nevertheless, continuous speech recognition systems are nowadays reasonably successful, although their performance of course strongly depends on the size of their vocabulary.

Word spotting can be regarded as a special form of isolated word recognition: the recogniser is "listening" for a limited number of words. These words may come embedded in background noise, possibly consisting of speech of competing speakers, or may come from the target speaker who is producing the word embedded in extraneous speech.

3.3.2.2.4 Corpora for speech recognition research

In general, two similar speech corpora are needed for the development of speech recognition systems: one for the training phase and one for the testing phase. The training material is used to set the model parameters of the recognition system. The testing material is used to determine the performance of the trained system. It is necessary to use different speech data for training and testing in order to get a fair evaluation of the system performance.

For speaker-dependent systems, obviously the same speaker is used for the training and testing phase. For speaker-independent systems, the corpora for training and testing could contain the same speakers (but different speech data), or they could contain different speakers to determine the system's robustness for new speakers.

When a system is designed for isolated word recognition, it should be trained

and tested with isolated words. And similarly, when a system is designed for telephone speech, it should be trained and tested with telephone speech. The design of corpora for speech recognition research thus strongly depends on the type of recognition system that one wants to develop. Several large corpora for isolated word (e.g. TIDIGITS) and continuous speech recognition (e.g. Resource Management, ATIS, BREF, EUROM, TIMIT and Wall Street Journal) have been collected and made available (cf. Appendix L).

3.3.2.3 Spoken language systems

Speech synthesis and speech recognition systems can be combined with natural language processing and Dialogue Management systems to form a Spoken Language System (SLS) that allows an interactive communication between man and machine. A spoken language system should be able to recognise a person's speech, interpret the sequence of words to obtain a meaning in terms of the application, and provide an appropriate response to the user.

Apart from speech corpora needed to design the speech synthesis and the speech recognition part of the spoken language system, speech corpora are also needed to model relevant features of spontaneous speech (pauses, hesitations, turn-taking behaviour, etc.) and to model dialogue structures for a proper man-machine interaction.

An excellent overview of spoken language systems and their problems is given in Cole (1995). The ATIS corpora mentioned above exemplify the type of corpus used for the development of SLS.

3.3.2.4 Speaker recognition/verification

The task of automatic speaker recognition is to determine the identity of a speaker by machine. Speaker recognition (usually called speaker *identification* can be divided into two categories: *closed-set* and *open-set* problems. The closed-set problem is to identify a speaker from a group of known speakers, whereas the open-set problem is to decide whether a speaker belongs to a group of known speakers. Speaker *verification* is a special case of the open-set problem and refers to the task of deciding whether a speaker is who he claims to be.

Speaker recognition can be *text-dependent* or it can be *text-independent*. In the former case the text in both the training phase and the testing phase is known, i.e. the system employs a sort of password procedure. One popular example of password-like phrases are the so-called "combination lock" phrases, consisting of sequences of numbers (mostly between 0 and 99) or digits (between 0 and 9). LDC provides a corpus for training and testing speaker verification systems based on combination lock phrases consisting of three numbers between 11 and 99 (e.g. 26–81–57) (Campbell 1995).

Knowledge of the text enables the use of systems which combine speech

and speaker recognition. In other words, before granting access to data or premises the speaker verification system can request that the claimant says the combination lock; the system then checks both the correctness of the password and the voice characteristics of the speaker. However, most password systems are susceptible to fraud using recordings of the passwords spoken by a customer. One way of making fraud with recordings much more difficult is by the use of *text prompted* techniques, whereby the customer is asked to repeat one or more sentences randomly drawn from a very large set of possible sentences. Again, the system checks both the contents of the reply and the voice characteristics of the speaker. Since surreptitious recording of millions of specific utterances is impossible, text prompted speaker verification systems should offer a very high level of security and immunity to fraud.

In the case of text-independent speaker verification the acceptance procedure should work for any text in both the training or the testing phase. Since this approach is especially susceptible to fraud by playing recordings of the customer, text-independent verification technology should best be combined with other fraud prevention techniques.

There are various application areas for speaker recognition, for instance helping to identify suspects in forensic cases, or controlling access to buildings or bank accounts. As with speech recognition, the corpora needed for speaker recognition or speaker verification are dependent on the specific application. If, for instance, the technology is based on combination lock phrases, a training database should obviously contain a large number of connected number or digit expressions. For the development of textindependent speaker technology there are no strict requirements as to what the training speakers say.

Corpora for the development and testing of *speaker* recognition systems differ in a crucial aspect from corpora collected to support *speech* recognition. For *speaker* recognition research it is absolutely essential that the corpus contains multiple recordings of the same speaker, made under different conditions. There is a range of conditions that should ideally be sampled, in order to be able to build a model of the natural variation in a person's speech due to realistic variations in the conditions under which the speech is produced.

Conditions to be sampled and to be represented in a corpus can be divided into two broad groups, viz. channel conditions, and physiological and psychological conditions of the speaker.

3.3.2.4.1 Channel conditions

The details of the acoustic speech patterns depend heavily on the acoustic background in which the speech was produced and on the response of the

transmission network. A corpus for speaker recognition research should at least include multiple recordings of the speakers made with different microphones or telephone handsets. Especially the transmission differences between carbon button and electret microphones in telephone handsets are known to affect the performance of speaker recognition systems. In this context, attention should also be paid to the different transmission characteristics of the fixed, landline telephone network and the rapidly growing cellular networks.

In actual practice, it is much more difficult to obtain a representative sampling of acoustic backgrounds. Yet one must be aware that loud background noise affects the speech in two ways, one of which is highly non-linear and therefore cannot easily be compensated for: in addition to decreasing the signal-to-noise ratio, background noise is likely to cause the speaker to change his speaking behaviour. The best known effect is known as the *Lombard effect*: in a high noise environment speakers tend to raise their voice level, and therewith their phonation style and probably also their articulation style.

3.3.2.4.2 Psychological and physiological conditions

The type of speaker variation addressed under this heading is also very difficult to sample. Given the practical limitations of a corpus collection project it is hardly feasible to require that each speaker be recorded in perfect health conditions, as well as when having a cold, the flu, or any other mild disease.

One simple approximation to sampling within speaker variation that is feasible from a practical point of view is to record speakers at different times of the day (early morning, noon, late night), and on different days of the week. In any case, the period over which the recordings are extended should span at least a couple of months. One might also consider recording speakers in completely sober conditions and after the consumption of a reasonable amount of intoxicating drugs.

More detailed accounts of speaker recognition can be found in O'Shaughnessy (1986) and Gish and Schmidt (1994).

Developing and testing speaker recognition systems with a database containing only a single recording session for the speakers should be avoided, because such databases cannot possibly account for even the slightest degree of within-speaker variation. Results reported on such databases (e.g. TIMIT) grossly overestimate the performance of the system being tested.

3.4 Specification of the linguistic content

3.4.1 Different types of speech data

The speech material in a corpus can vary from isolated sounds to complete conversations. In general, the extent to which the experimenter has control over the speech material decreases as it becomes more and more spontaneous and natural. The term *natural* refers to a rather intuitive concept that can be interpreted in different ways. We regard speech to be maximally natural when two or more speakers have a conversation in a familiar environment about a subject they themselves choose to talk about, since this is presumably the situation for which speech was "invented". Although read aloud speech is a commonly used speaking style (and may be regarded as a natural speaking style from a sociolinguistic point of view), we regard this style as derived from the most natural style mentioned above. When reading a text, people have the tendency to speak more formally and to articulate more carefully than when they are involved in free conversation. Thus, in our opinion the naturalness of speech should be judged on a gradual scale. It should be noted that *control* over the speech material is not always necessary and may even be counterproductive, especially when one wants to study the variation of speech as a function of communicative context. However, strict control over the speech material is required for some applications, such as the development of speech synthesis systems. In the following, eight types of speech data will be distinguished.

3.4.1.1 Read aloud isolated phonemes

Vowels pronounced in isolation (or in a "neutral" context, such as /hVt/) are often used as the frame of reference for experiments in which vowels from connected speech are investigated. Continuant consonants, e.g. /l, r, w, j, n, m, s, f/, can also be pronounced in isolation. Non-continuants, e.g. /p, t, k, b, d, g/, must be followed or preceded by a vowel, e.g. the "neutral" schwa.

3.4.1.2 Read aloud isolated words

Isolated words can be either "nonsense" words or existing words. In the case of nonsense words the experimenter can create all possible kinds of phonotactically correct sound sequences. This gives the opportunity to study coarticulation in a systematic way. Nonsense words are also used to extract models for a dictionary of phonetic elements when a synthesis system is developed. When existing words are used, the number of possible sound sequences is restricted to what is phonotactically appropriate in the lexicon of a given language. It must be realised that control over the sounds produced by the speakers may not be perfect, because the pronunciation of

polysyllabic words can be influenced by the stress pattern, which may be ambiguous (cf. words like *record* in English).

When speakers have to read aloud a list of isolated words, their pronunciation may be influenced by the orthographic representation of the words, a phenomenon known as *spelling pronunciation*. Spelling pronunciation is especially apparent in languages which form nominal compounds; if sound sequences occur across the morpheme boundaries that would lead to assimilation and degemination in connected speech, one should still anticipate that in reading aloud all sounds are realised. This phenomenon can be circumvented by having the speakers name the words through the presentation of pictures, but this method can only be applied to a limited number of words. It is, for instance, not suitable for abstract concepts.

3.4.1.3 Read aloud isolated sentences

The carrier sentence is one type of an isolated sentence. Carrier sentences are often used when one wants to get a somewhat more natural pronunciation of (nonsense) words in comparison with words spoken in isolation, especially with respect to speech rate. The test words are embedded in the carrier sentence, as illustrated by the example "I will say -a test word -a again". The same carrier sentence is used repeatedly for all occurring test words, so that the influence of the acoustic and linguistic context on the test words is controlled.

More natural speech material can be obtained when "normal" (linguistically meaningful) sentences are constructed by the experimenter. Such sentences can be used to train phoneme based recognisers and to study, for instance, word stress or coarticulation in a relatively natural linguistic context. One should note that an isolated sentence may be interpreted by a speaker in a wider semantic context, which can influence the pronunciation of the sentence, especially with respect to the position of sentence accent(s). Sometimes a semantic relation between subsequent "isolated" sentences may arise as a result of the specific ordering of the speech material. Since individual speakers may imagine different semantic contexts for a specific sentence, variability in the suprasegmental features of the test sentences can occur. If desired, this variability can be reduced by using punctuation and other typographical means (for instance, capitals or boldface characters) to indicate words that should have a sentence accent. A more natural way of doing this is to let each sentence be preceded by a question that evokes sentence accents at the desired positions. It should be clear, however, that neither practice can be recommended in the collection of large corpora of telephone speech.

For many purposes, such as the development of a phoneme-based recogniser, it is crucial that all phonemes are represented in the speech corpus in sufficiently high numbers. Due to the large differences in frequency of occurrence of the phonemes in the language in general, uniform phoneme frequencies will not obtain in randomly chosen sentence material: such material will, instead, reflect the differences in phoneme frequencies. It is proposed to reserve the term *phonetically balanced* for speech material containing phonemes according to their frequency of occurrence in the general language. *Phonetically balanced* sentences may be used for speech audiometry and for testing the transmission characteristics of communication channels or public address systems.

Approximately uniform phoneme frequency distributions can be achieved by using *phonetically rich* sentences. For that purpose greedy algorithms (Van Santen 1992) can be used. Suppose you want to have a set of sentences in which each phoneme of the language of interest occurs at least once. Of course, you could try to create this set of sentences yourself, but this would be difficult and time-consuming. Furthermore, you might end up with sentences that look rather "constructed". An alternative would be to search for an appropriate set of sentences in a sufficient large text corpus, for instance, a large amount of newspaper data on CD-ROM. An advantage of this procedure is that much more variation in the sentences is obtained. A greedy algorithm can be used to obtain the minimum number of sentences containing all phonemes. The following steps have to be taken to get the desired test set:

- 1. Use a grapheme-to-phoneme converter in order to be able to search for phonemes instead of characters.
- 2. Select the sentence in the corpus with the largest number of phonemes, not counting phonemes that are repeated within the same sentence.
- 3. Select each next sentence as the one with the largest number of phonemes that have not yet been covered. Stop this procedure when the entire set of phonemes has been covered.

To obtain more occurrences of each phoneme, the procedure described above can be repeated any number of times with the remaining sentences in the text corpus. Of course, the greedy algorithm can also be used for other basic units than phonemes, for instance: characters, diphones, vowels in specific consonantal contexts, subsets of words, or specific discourse units. The greedy algorithm can be amended in various ways. For example, one can maximise coverage of high frequency units by using the sum of the frequencies of the units in a sentence as criterion. This may be important when complete coverage of all units is impossible, in which case one likes to cover the most frequent units first. Furthermore, all kinds of constraints can be imposed on the sentences that are selected, for instance with respect to their length, word material, syntactic structure, etc. Note that you can also choose other contexts for the basic units than sentences, such as clauses, or words. For example, you might want to search the text corpus for the minimal set of words in which each phoneme occurs at least once.

It should be clear that very large text corpora may be needed to obtain a sufficient number of phonetically rich sentences. In some corpora phonetically rich does not only mean that an attempt has been made to obtain uniform phoneme frequencies, but also uniform diphone or triphone frequencies. When designing phonetically rich sentences intended to be read by members of the general public (e.g. in the POLYPHONE corpora) care must be taken to avoid very long sentences, because these are extremely difficult to read aloud. Moreover, all sentences must be checked for very rare words (which are likely to cause reading problems) and for contents which are potentially insulting. In POLYPHONE, candidate sentences had to contain at least four words and a maximum of 80 characters, including spaces and punctuation marks. The latter condition guarantees that the prompting text will not span more than two lines on the prompting sheet (40 characters per line). Selection of sentences on the basis of length and phonemic contents can be done automatically. However, checking for insulting contents must be done manually.

3.4.1.4 Read aloud text fragments

The naturalness of the produced speech may increase even more when speakers read aloud a series of sentences that are semantically related, *provided that the subject is able and used to reading aloud paragraph length material.* The prompting material can consist of a text fragment taken from, for instance, a newspaper or a book (e.g. BREF, Wall Street Journal). But the text fragment can also be created by the experimenter, when it is necessary to impose some specific restrictions on the speech material, for instance with respect to phonemic structure, word structure, or syntactic structure. Reading aloud a text fragment is more difficult than reading aloud a list of isolated sentences. It is very likely that the speech produced by different speakers who are asked to read a text fragment will vary considerably, especially with regard to aspects like vividness, speech rate, omitted speech segments, prosody, etc. The preferred position of sentence accents in a text fragment can be indicated with capitals or boldface characters. This is not recommended if one is interested in more natural speech.

3.4.1.5 Semi-spontaneous speech

When speech corpora are gathered for commercial applications, a common task of speakers is to read numbers or alpha-numerical expressions, such as ZIP-codes. Speakers have to some extent the freedom to pronounce these numbers or alpha-numerical expressions as they like. For example, there appear to be substantial differences between the ways in which subjects express telephone numbers. Some may read the telephone number as a string of digits, whereas others may read it as a string of numbers containing two or more digits. In addition, it may make a difference whether the telephone number is familiar (for instance, a friend's number), or unfamiliar. The POLYPHONE corpora are good examples of corpora that contain such semi-spontaneous speech.

3.4.1.6 Spontaneous speech about a predetermined subject

The previous types of speech material were all concerned with the *reading aloud* of some piece of text by *one* speaker at the time (disregarding the naming of words through the presentation of pictures). In the present section we will discuss *spontaneous* speech from *one or more* speakers. The major difference between read speech and spontaneous speech is that the former fixes vocabulary and syntax, whereas the latter leaves speakers free to choose their own vocabulary and syntax. The naturalness of the produced speech increases when speakers are allowed to choose their own words. In order to keep some control over the speech material, the experimenter can determine the subject the speaker has to talk about.

The subject of conversation is relatively fixed when speakers are asked to retell a story that they heard or read shortly before. Since it is likely that speakers will use at least some of the words that occurred in the story, this method allows the experimenter to gather "spontaneously" spoken versions of specific words of interest. In a variant of this method, speakers can be asked to invent a story based on a cartoon (without text balloons), or on some complex picture that is bound to evoke the words of interest. In all these designs, monologues are involved, although a session manager may try to guide the discourse in the desired direction. However, one should be aware that many naive subjects do not feel at ease in a situation in which they must maintain a monologue for an extended period of time. Most people feel much more comfortable in a dialogue situation. Moreover, interview situations provide some additional control over subjects' speech, because the interviewer determines the subject of conversation, and subsequently guides the conversation in the desired direction.

Another kind of guided spontaneous speech is an information dialogue: people who attempt to obtain information about, for instance, train or plane schedules. Speakers ask information from an information agent or a computer system about time and place of departure, destination, etc. In this way spontaneous speech can be obtained, even if it concerns a very restricted subject. This paradigm is used in the (D)ARPA Air Line Travel Information System (ATIS) task. Train time table information dialogues are now being recorded in several languages, e.g. German, Dutch, Italian, French, British English etc. in the MLAP projects MAIS and RAILTEL.

Although a speech situation with two or more people is more natural than a monologue, overlapping acoustic material may result from several people speaking simultaneously. For some applications, such as research on basic speech processes, overlapping acoustic material is difficult or impossible to

use. Of course, one can try to extract speech fragments from recorded dialogues in which only a single speaker is talking. The study of simultaneous speech from two or more speakers is important for research on dialogue or discourse analysis, intention analysis, and spoken language understanding. The gathering of multiple simultaneous speaker corpora is still in its infancy. Such corpora are indispensable for studying speech in all its relevant aspects. In addition, speech recognisers, which are up to now only able to deal with one speaker at a time, would eventually also have to be able to deal with different speakers talking simultaneously. Speech corpora containing dialogues could supply the training and testing data for such advanced recognisers. To make such corpora useful for research and development purposes each individual speaker should be recorded on a separate track, using a microphone array with very high directional sensitivity. Additional tracks can then be synthesised, simulating less perfect directional sensitivity. Alternatively, subjects could be recorded in a teleconference, although such distributed recordings would require extensive precautions to allow one to synchronise the tracks originating from completely independent recorders. A special type of information seeking dialogue, which is becoming increasingly important, is the one between a human and a computer. In order to gain a clear insight into the way people behave when they have to interact with computers, in the absence of computers that can entertain such a conversation, the Wizard of Oz technique was invented. This technique will be briefly described in the next section.

3.4.1.7 The Wizard of Oz technique

In the children's novel The Wizard of Oz (Baum 1900) a young boy is bullied by an oracle called the Wizard of Oz. The crux of the story is that the Wizard of Oz turns out to be nothing more than a device operated by a man. In the Wizard of Oz technique a human plays the role of the computer in a simulated human-computer interaction. Of course, the easiest way to learn about the way humans behave when they have to interact with computers would be to actually have them interact with a computer. However, in order to be able to build a computer system that can participate in a dialogue with a human, one has to know how a human-computer interaction is likely to proceed. The Wizard of Oz technique can be seen as an intermediate step in the design of such a computer system. Because the subjects who participate in a Wizard of Oz experiment have to be convinced that they are actually talking to a computer, some precautions must be taken. For example, the wizard simulating the computer should be talking with a "computer voice" (in the case that spoken output is required), and the wizard should also make deliberate errors similar to the ones that a computer could be expected to make in the application of interest.

As spoken language systems are rapidly approaching a performance level

that is acceptable for an increasing range of applications, it seems likely that man-machine dialogue systems will be used more and more in the near future. For the development of such systems speech data gathered in Wizard of Oz experiments will be indispensable, as long as at least one part of the system is not yet good enough for experiments with large groups of users. A more comprehensive discussion of the Wizard of Oz technique is given in Chapter 13.

As the performance of SLSs improves, the development of new applications will be increasingly based on pilot experiments with *a system in the loop*, i.e. with test versions of the application in which the wizard is replaced by a computer system which has enough functionality to support the man-machine interaction.

3.4.1.8 Spontaneous speech

This type of speech material, in which speakers are allowed to freely choose their own words and their own subject of conversation is most natural, especially in a dialogue situation. Most remarks that were made in the previous sections, also apply to the present one. As with all natural processes, the observer's paradox can play a role in the recording of spontaneous speech: in order to obtain speech that is as natural as possible, the researcher has to observe how people speak when they are not being observed (Labov 1972). To overcome this methodological paradox, several techniques have been proposed throughout the development of sociolinguistic research (Argente 1991):

• TRYING TO GIVE SUBJECTS THE IMPRESSION THAT OBSERVATION IS AB-SENT

Subjects are usually unaware of being observed when they are (surreptitiously) recorded before or after the actual experimental procedure (reading sessions, interviews, or whatever). Also (planned) interrupting phone calls, or interruption by a third person during the recording session can give a subject the impression that observation has temporarily stopped. ASKING SUBJECTS ABOUT EMOTIONAL EVENTS IN THEIR LIFE

- The investigator can, for instance, ask subjects to remember occasions when they were in physical danger. Speech data available from radio or television interviews with witnesses of an actual disaster (a plane crash, an earthquake, etc.) could contain realistic samples of emotional speech. This technique is based on the belief that emotional involvement triggers unattended forms of speech.
- Using a member of the subject's own community for interviews or conversations

This technique is based on the belief that people tend to talk in a more casual way to people with a similar social or geographical background. Especially recordings of people who are friends in everyday life are likely to contain samples of the vernacular.

• Focussing research on adolescents

This strategy is based on the belief that adolescents have not yet developed the full-fledged range of speech behaviour and communicative competence characteristic of sophisticated adults and that adolescents therefore may be inclined to use a more casual speech style even in the presence of a recording device.

3.4.2 Factorial experiments and corpus studies

Experimental speech research has traditionally been focussed on factorial experiments, that is, experiments in which a number of factors are defined that are hypothesised to influence some aspects of speech behaviour, in production or in perception (see Chapter 9). The amount of speech in these experiments has typically been small, if only because it was practically impossible to record large amounts of speech in production experiments or to generate large amounts for perception experiments. The major causes of the limitations were in the tight control of the speech needed for well designed factorial experiments and in the time required from the subjects. Tight control is necessary to prevent the outcome of factorial experiments from being meaningless: this type of experiment requires that all conceivable factors different from the small number under study be kept constant, whereas the experimental factors are varied over a limited range. It is not our intention to criticise factorial experiments, if only because they have contributed to virtually all the knowledge we have about speech and because until recently there was hardly an alternative. But it must be acknowledged that, precisely because of the tight control, the speech used in the older experiments may not have been exactly "communicative". In the majority of the cases the subjects performed in situations which are quite remote from normal communicative behaviour; therefore, some caution should be exercised in generalising the results of controlled experiments to "normal communicative" speech.

Another reason to be careful in interpreting results of factorial experiments is the possibility that the experimenter did not completely succeed in keeping all non-experimental factors constant: it may be the case that nonexperimental factors did co-vary with experimental ones, thereby being responsible for at least part of the effects attributed to the experimental factor(s). One case in point is intonation research, that has been pretty much focussed on pitch and on duration effects. There is, however, increasing evidence that other factors like spectral structure, spectral slope, spectral dynamics, etc. also play a role, and perhaps one that is quite important. In short: there is a danger that factorial experiments lead to overestimating the impact of the factors under investigation, at the cost of factors that were supposed to be constant, but that actually co-varied so as to enforce the effects of the experimental factors.

Now that very large corpora are becoming available, it is possible to set up

another type of experiment, in which the behaviour of one or more specific factors is investigated in a very large, perhaps comprehensive number of different contexts. Instead of trying to neutralise the effect of concomitant factors by trying to keep them constant (which will normally mean that one of the many different levels of such factors is selected, e.g. a voiceless stop as the right neighbour of the phonemes under study, or only syllables which have a prominence lending High-Low pitch contour), one may try instead to sample many different contexts. Of course, in order to make this type of research feasible, one has to assume that subject effects can be treated in exactly the same way as context effects, because it will still be extremely difficult to have subjects perform for very long periods of time. In designing corpus based experiments one must be aware of the extreme skewing of many frequency distributions observed in spoken language. For instance, in all languages for which data on phoneme frequencies are available it has appeared that within a system some phonemes occur much more often than other phonemes. Random sampling would leave one with a very high likelihood of missing infrequent phonemes and of missing possible contexts, unless the total corpus is made excessively large. Greedy algorithms (cf. Van Santen 1992) can be used to find the minimum amount of linguistic material that covers a maximum number of phenomena, but even with the use of greedy algorithms it cannot be guaranteed that all possibly relevant conditions are indeed covered: conditions which are not formulated as targets for the search will only be present by chance. Since complete coverage is not practically attainable, corpus research must deal with missing data in one way or another. Attempts have been made to handle missing data by means of knowledge-based arithmetic models, including all relevant parameters; alternatively, "blind" statistical modelling techniques like CART (Classification And Regression Trees) can be used. There seems to be some preference for arithmetic models, unless one can guarantee that the missing data are not concentrated in a few subspaces (cf. Van Santen 1994).

3.5 Specification of number and type of speakers

In addition to specifying the linguistic content of a corpus, the number and type of speakers in a corpus is the second major factor in specifying and classifying corpora. Due to their idiosyncratic characteristics speakers add substantially to the amount of variation present in a corpus.

3.5.1 Corpus size in terms of speakers

The number of speakers who are represented is one of the most important characteristics of a spoken language corpus. Based on the *number* of speakers in the corpus, speech corpora can be roughly divided into the following three classes:

- 1. speech corpora with few speakers,
- 2. speech corpora with many (about 5 to 50) speakers,
- 3. speech corpora with very many (more than 50) speakers.

3.5.1.1 Speech corpora with few speakers

Such corpora are often used in the development of speech synthesis systems. In most cases the speech of one or two persons (typically one man and one woman) is recorded. The corpus is used to prepare dictionaries of phonetic elements (allophones, diphones, etc.), and to design prosodic models. The speech material may consist of nonsense words in which sequences of phonetic elements are systematically varied, and a series of sentences to extract prosodic rules. For developing synthesis systems it is recommended to use experienced speakers. Especially when recording the material that serves for building the segment inventory it is extremely important that the speakers can keep pitch, loudness, voice quality and tempo constant.

Corpora comprising very few speakers are also common in basic speech research, especially where invasive measurements must be made. Corpora in this domain typically contain several additional signals recorded simultaneously with the acoustic speech signal, see e.g. the ESPRIT Basic Research Project on Articulatory Phonetics. The additional signals can range from the Electroglottogram (which was also recorded in part of the EUROM-1 corpus and the Transnational English Corpus) to subglottal pressure recorded via tracheal puncture and EMG activity of intrinsic laryngeal muscles. It should be emphasised that very few speakers does not necessarily imply a *small corpus*. For instance, when one needs to record one speaker producing all three-consonant clusters in languages like Dutch, English or German, in all possible phonetic contexts, within syllables, across syllable boundaries, across word boundaries, in stressed and unstressed syllables, at several positions in a sentence or in a prosodic contour, the amount of speech required is formidable, even when greedy search algorithms (cf. Van Santen 1992) are used to find the smallest possible number of sentences which comprise all contexts.

Similar remarks apply to intonation and prosody in general. If a textto-speech system is developed that must be employed in many different applications (reading factual information in e.g. a train time table information system, or reading popular daily newspapers to blind subscribers), enormous amounts of speech are needed to capture all relevant prosodic phenomena.

3.5.1.2 Speech corpora with about 5 to 50 speakers

Speech corpora of this size are often used in experimental factorial research. The speech material can range from isolated nonsense words to a complete discourse, dependent on the specific application. Studies on prosody, for instance, would require linguistic units that exceed the word level. Speakers can be men, women, or children. The speech can be recorded under high quality laboratory conditions, but also "in the field". In general, the number of speakers and the number of repetitions of the speech phenomena that are investigated should be large enough for a meaningful statistical processing if factorial experimental designs are planned. The power of a statistical test depends on the number of independent observations. If a corpus is developed for a factorial experiment, standard procedures are available and should be adhered to for determining the minimum number of speakers and/or the minimum number of utterances per speaker to allow planned statistical tests to reach a pre-specified power. These standard procedures can be found in most textbooks on statistics, such as Hayes (1963) (pp. 269–280), Ferguson (1976), or Marascuilo and Serlin (1988). In designing very large vocabulary speech recognition systems, on the other hand, one will strive for a maximally broad coverage of relevant phenomena, probably at the expense of high numbers of exact replications of specific (relatively rare) phenomena Chapter 9 should be consulted on methodology.

3.5.1.3 Speech corpora with more than 50 speakers

Speech corpora of this composition are necessary to adequately train and test speaker-independent recognition systems. Speakers can be men, women, or children, dependent on the application. The speech material can be limited to a list of isolated words or numbers, but it can also contain read aloud sentences and paragraphs or even spontaneous speech in the case of interactive dialogues. Speech may be recorded under laboratory conditions or in (quiet) offices, but if a telephone recognition system is involved, the speech corpus should, of course, consist of telephone speech both for the training phase and the testing phase.

3.5.1.4 General remarks

Of course, possible applications of corpora may be quite different from the typical ones listed above; some fundamental research may, for instance, require a very large speech corpus, whereas a simple recognition system may be developed with a rather small speech corpus. Furthermore, the list of applications of speech corpora given above is not meant to be exhaustive, but it should help to illustrate the large differences between speech corpora, depending on their research goal. *Speaker Recognition* is a branch of speech (technology) research which has received little attention in the past decade. This is reflected in the lack of publicly available corpora to support speaker recognition research. However, it is not completely true to say that there are no corpora which are suitable for speaker recognition research; notable exceptions are the King corpus and the Switchboard corpus, both available though the LDC (cf. also the Proceedings of the ESCA workshop on Speaker

Recognition in Martigny, April 1994). For a corpus to be suitable for speaker recognition research it is essential that speakers are recorded more than once, and that recordings are made at different days, in different realistic acoustic environments and with different microphones.

3.5.2 Speaker characteristics

How are the speakers for a speech corpus selected? Again, this strongly depends on the application one has in mind. For the development of a speech synthesis system, experienced speakers, such as news readers or actors, are most appropriate. For the training and testing of recognition systems, on the other hand, the population of interest must be suitably sampled. There is no general agreement on the exact meaning of "suitable" in this context. One definition would amount to random sampling of the population of interest. This operationalisation usually results in different numbers of samples from subpopulations in the population of interest. For example, when the total population of army personnel is sampled, the subpopulation of women is likely to be poorly represented. In the case of the training and testing of a recognition system for the army, this female under-representation might seem to be acceptable, because the recogniser would have to deal mainly with male speakers. However, it may appear that some of the influential heavy duty users are women and then the recogniser should better be designed to handle the few but important women with the same performance as for men. In general, random sampling has the potential drawback that extremely large numbers of samples are needed to ensure that rare, but nevertheless important phenomena are included. When, where, and why rare phenomena may still be important depends on the application for which the corpus is collected. In the case of fundamental research, on the other hand, the aim is often to compare subpopulations in some respect, and then it would be more appropriate to draw an equal number of samples from all subpopulations of interest. Uniform sampling of all subpopulations of interest ensures that all relevant variation is included in the corpus with the smallest possible number of speakers. The application for which the speech corpus is collected not only determines the best sampling strategy, but it also influences the choice of speakers. For example, speech processing often involves spectral analysis of the recorded speech. Several analysis techniques, such as pitch extraction or formant extraction, are less accurate for high-pitched voices (women and children) than for low-pitched voices (men). If such analysis techniques are used and the sex of the speakers is of no concern for the research goal, it would thus be sensible to select only men for the speech corpus. In general, however, it is recommended to include all possible types of speakers in a speech corpus, unless there are imperative arguments to exclude specific speaker groups. Specifically, it is strongly recommended to include equal numbers of females and males in

each corpus. Speaker characteristics, which are potentially important and should therefore be considered when selecting the speaker population are described and discussed below.

3.5.2.1 Stable / transient speaker characteristics

The many speaker characteristics that may influence the speech signal can be divided in two main classes: relatively stable characteristics, and transient (temporary) characteristics. Stable speaker characteristics comprise on the one hand physiological and anatomical factors such as sex, age, weight, height, smoking/drinking habits, and possible pathologies, and on the other hand geographical and sociolinguistic factors. Transient (temporary) speaker characteristics cover factors such as a cold, or other mild afflictions of the speech organs, general physical condition (dependent on, for instance, the number of hours of sleep during the previous night), stress, and emotional state. Whereas transient speaker characteristics are very difficult to control, stable speaker characteristics are easier to take into account in the design of the speech corpus. For an overview of several important stable speaker characteristics, we refer to Scherer and Giles (1979). The most important stable speaker characteristics will be mentioned below.

3.5.2.2 Demographic coverage

Demographic factors form a very important set of relatively stable speaker characteristics which must be considered when designing sampling procedures for a corpus collection project. Each corpus should have sufficient demographic coverage. However, it is not always possible to determine all potentially relevant demographic factors *a priori*. Nor is the distribution of all factors in the total population always known. It is likely that the availability of detailed and reliable demographic data differs between the European countries. The availability such data in less developed countries is even more questionable. In selecting speakers for inclusion in a corpus the possibility to assess certain characteristics is dependent on the recording protocol. If randomly selected speakers are recorded over the telephone, many personal characteristics cannot reliably be collected: self-report from the speaker is the only means of gathering the data.

3.5.2.3 Male / female speakers

Sex (gender) distinctions are known to have an enormous impact on speech quality. It is not well known at what age sex-related speech characteristics become prevalent. There is some evidence that sex-related speech characteristics are only partly due to physiological and anatomical differences between the sexes; cultural factors and sex role stereotypes also play an important role. Therefore, it is possible that the age at which sex-related

differences become apparent differs between cultures and therefore between languages. See for general information on sex-related speech characteristics Smith (1979), Coates (1986), Philips et al. (1987), and Brouwer and De Haan (1987). For the time being, no definitive recommendations can be given with respect to the age above which sexes should be distinguished and sampled individually. Unless the contrary can be motivated from the specific application the corpus is collected for, each corpus should comprise approximately equal numbers of speakers of both sexes. For some applications, recordings of young children may also be required. Children should be considered as a "third sex", independent of adolescent or adult females and males. Speaker sex is known or suspected to affect at least four aspects of speech behaviour.

1. PITCH AND INTENSITY

Women are known to have higher average pitch than men. There are also indications that average intensity in female voices is somewhat lower than in male voices. In particular, higher pitch may affect spectral analysis techniques: pitch and formant extraction may be less accurate for high-pitched female voices than for low-pitched male voices. When a corpus is recorded to develop and test parameter extraction techniques, a realistic proportion of high-pitched female voices should be present. It should be realised that there is an interaction between sampling rate and the accuracy with which pitch frequency can be determined. In female and child speech even 20 kHz sampling frequency may not be high enough to obtain sufficient accuracy, as pitch frequencies in especially child speech may be as high as 500 to 750 Hz. Fortunately, sampling frequency can be increased using straightforward signal processing procedures whenever the need arises.

2. Overall spectral slope

Women are reported to tend more towards a breathy voice quality than males. It is not known whether this tendency is related to physiological and anatomical factors or whether it is mainly due to culturally determined sex role stereotypes. Overall steeper spectral slope causes problems for some parametric signal processing techniques (e.g. formant extraction).

3. Accuracy of pronunciation

Women are reported to adhere more to standard pronunciation than men (Labov 1972). It is not known whether this finding generalises to all languages. It remains to be seen whether sex related pronunciation variation is best modeled and described on the level of phonemic representations of words or on the level of the phonetic implementation of what is essentially the same phonemic form. Awaiting results of further research in additional languages/cultures this factor is probably not sufficiently important to attribute great importance to it. Moreover, this aspect is very difficult, if not impossible, to separate from other sex-related factors, and will therefore be duly represented as long as the sexes are adequately represented in the corpus. Variation in pronunciation accuracy may also be caused by factors related to age and social status.

4. Vocabulary and syntax

Sex-related differences in vocabulary and syntax are certainly culturally determined. Here, the factor sex interacts with factors like age and social status. Differences on the level of vocabulary and syntax are only relevant when spontaneous speech is being recorded. If all speech material consists of read utterances, vocabulary and syntax are completely determined by the prompting material. However, the ability to pronounce can depend on socio-economic status or education.

3.5.2.4 Age

Although the impact of speaker age on speech behaviour has not received much attention in previous research, there are indications that age influences at least two aspects of speech behaviour (Helfrich 1979).

3.5.2.4.1 Voice quality

There has been some research on the relation between age and voice quality. Most studies were concerned with the question whether speaker age can reliably be estimated from the speech signal alone. It seems that people are moderately good at guessing age from speech signal characteristics, although reported correlation coefficients may be mainly determined by the ability to discriminate between very young, very old and adult but nonsenior groups. The exact signal characteristics which enable people to guess the speaker's age are not well understood; neither is it possible to estimate their impact on the performance of automatic speech and speaker recognition. Until the questions about the importance and the exact nature of the impact of age on speech signals have been answered, it is recommended that attempts be made to sample the relevant age groups. In doing so, a distinction should be made between the group under 20, the group between 20 and 60 and the group over 60. If relevant, the group under 20 should, of course, be subdivided into toddlers, children, adolescents and young adults. However, the exact ages separating these subgroups is the subject of discussion. Moreover, in many respects mental and physiological maturation may be more important than calendar age.

3.5.2.4.2 Vocabulary and syntax

Here the considerations described above in the paragraph on the impact of sex on speech behaviour apply in exactly the same way. There is some literature suggesting that vocabulary and syntax of the older generation are different from the younger speakers, but apart from obvious observations that the subjects spontaneously discussed by senior citizens tend to differ there is little hard data to support the claim that age is more important a factor than, for instance, social group and education level.

3.5.2.5 Weight and height

As with speaker age, most research in the past has concentrated on the question whether people can estimate speaker weight or speaker height from speech recordings alone (Van Dommelen 1993). It appears that people are moderately successful in this task. It will be clear that weight and height of speakers are highly correlated. The exact signal characteristics that enable people to guess the speaker's weight and height are not known. In a sufficiently large sample of speakers, most weight/height groups will probably be represented.

3.5.2.6 Smoking and drinking habits

Several investigations have shown that voice quality can change under the influence of smoking or the use of alcohol (Gilbert and Weismer 1974). One of the most common consequences of smoking and drinking is premature ageing of the mucous membrane covering the vocalis muscle, resulting in a hoarse voice quality. Excessive drinking may eventually result in brain damage, which may in turn lead to severe speech disorders. The use of drugs can have a similar effect. In those cases it would be more appropriate to speak of pathological speech.

3.5.2.7 Pathological speech

The boundary that divides pathological speech from non-pathological speech is very difficult to draw. Hoarseness due to smoking can be regarded as a very mild speech disorder, whereas more severe speech disorders include, for instance, paralysis of the vocal cords and aphasia. Speech disorders can be divided into two main classes: those where there is a clear organic (anatomical, physiological, neurological) cause, and those where there is not. The latter category is usually referred to as functional disorder. However, in many cases there is no clear cut distinction between organic and functional speech disorders; often both types are involved, or it is unclear which of the two types is involved. Speech disorders can be described at five different levels:

1. Articulation disorders

This involves the distortion, deletion, or substitution of sounds or sound combinations. Usually such disorders are functional, but they may also result from lesions of the lips (e.g., a cleft lip), the palate (a cleft palate), the teeth, the tongue, the jaw, or the nose. Another possible cause of articulatory disorders is dysarthria, a damage to the central or peripheral nervous system, manifested by neuromuscular disability.

2. Resonance disorders

This involves lesions of the oral, nasal, or laryngeal cavities. Apart from functional causes, resonance disorders can result from, for instance, surgical removal of the tonsils, a cleft palate, or nose polyps. 3. Voice disorders

This involves lesions of the vocal cords, referred to as dysphonia. The voice may emerge as a whisper (no vocal-cord vibration), for instance due to paralysis; or vocal-cord vibration may be present to some degree, but accompanied by excessive air flow (a "breathy" voice); or there may be irregular and therefore aperiodic vocal fold vibration, for instance due to the growth of abnormal tissue (nodules) on the vocal folds, resulting in a "hoarse" voice quality. Dysphonia may be caused by psychological and emotional factors, such as a severe shock, or by organic factors. A serious voice disorder is cancer of the vocal cords, which may lead to the surgical removal of the larynx (laryngectomy). Although the patients can learn alternative voicing mechanisms, their speech is usually severely degraded.

4. Language disorders

This involves disorders that do not affect the production of the speech message, but rather its content. These disorders are usually classified under the name *aphasia*. Patients suffering from aphasia may, for instance, use a reduced and incomplete sentence structure, have difficulty in wordfinding, use an inappropriate intonation, or make erratic pauses. The cause of aphasia is brain damage due to, for instance, a stroke, thrombosis, a tumour, an accident, or excessive drinking.

5. Rhythm disorders

The usual terms to describe the main rhythm disorders are stuttering (or stammering) and cluttering. Stuttering is a very complex phenomenon that is characterised by, for instance, a repetition of speech segments, abnormal prolongations of sound segments, words being unfinished, or circumlocutions to avoid types of sound that cause problems. Stuttering varies enormously from person to person and from situation to situation. It is, for instance, well known that stutterers almost never stutter when they are singing. Both organic (genetic) causes and functional (environmental) causes are assumed to underlie the stuttering phenomenon. Another major category of nonfluency is cluttering. The primary characteristic here is that the patient tries to talk too quickly, and as a result introduces distortions into his rhythm and articulation. The description and theoretical study of cluttering is less advanced than that of stuttering. In addition, there is a considerable overlap between the categories of stuttering and cluttering.

For many purposes it is most appropriate to build speech corpora with a large variety of speakers. However, the speaker variability should be kept within reasonable bounds. Severely pathological speech will, in general, deviate substantially from "normal" speech and thus it is usually not desirable to include this type of speech in a normal speech corpus. On the other hand, speakers with mild pathological disorders, such as hoarseness, can be included in for instance speech corpora designed for recognition.

Of course, research might focus specifically on pathological speech, for instance when a recogniser is developed for use as an environmental control device for handicapped persons. In that case pathological speech should
of course be amply represented in the speech corpus. Pathological speech should also be present in a corpus designed to cover as much speaker variation as possible (a kind of "all-purpose" speech corpus). A more elaborate discussion of pathological speech can be found in Perkins (1977) and Crystal (1980).

3.5.2.8 Professional vs. untrained speakers

Professional speakers should be selected when recording very large corpora with very few speakers, for instance to develop text-to-speech systems. The major reason to prefer professional speakers for this purpose is their ability to keep pitch, intensity and speech rate constant, not only during one recording session, but also over several sessions, which may have to be scheduled on different days, perhaps spread over several weeks or even months. One possibly important drawback of using professional speakers must be emphasised: more often than not, professional speakers are not really representative of the "normal" speech behaviour in the community. If the corpus is collected for the development of a text-to-speech system this may not be a problem. However, linguistic and phonetic findings based on a corpus comprising only speech of a small number of highly trained professional speakers should not be generalised without extreme caution.

3.5.2.9 Geographical and sociolinguistic factors

It is well known that both the regional and the sociolinguistic background of speakers can have a large effect on their speech. People speak differently depending on the specific region(s) in which they were brought up, and dependent on factors such as the linguistic background of the parents, social status, and education level. It is widely assumed that the high-school period is most decisive for the regional or dialectal colouring in one's speech. Therefore it is strongly recommended to obtain information about the highschool period when collecting data about the speaker's background.

Dialectal speech or regional/dialectal colouring of the prestige variant of a language, like Received Pronunciation (RP) in British English or *Hochdeutsch* in Germany, are known to be perhaps the most important source of speaker-related variation. Not all languages have a widely accepted and well documented pronunciation standard, like RP in English. Given the enormous amount of literature on Dialectology one would assume that the impact of dialects on standard speech is well understood. Unfortunately, this does not appear to be the case. Linguists and dialectologists appear to disagree about the number of major dialects in a language area, and about the boundaries between the areas where a specific dialect is spoken. Moreover, the majority of the dialect studies were based on written questionnaires. Although there are large amounts of recorded dialectal speech stored in the national Dialectology institutes, these recordings do not qualify as corpora, because they exist only on analogue tapes, with little or no detailed annotation. In collecting new corpora the factor regional/dialectal colouring should be properly accounted for. However, since the basic data to determine number of dialects and dialect boundaries are difficult to obtain and probably not always reliable, it is recommended that dialect is operationalised by geographic region. If necessary, processing of the corpus data can yield *post hoc* data on dialect differences. However, it has appeared that *post hoc* determination of the dialect background of a speaker as part of the transliteration/transcription process poses big difficulties. There is one additional factor which complicates the procedures for sampling dialectal influence, viz. the increasing mobility of the population. It is acknowledged that the impact of mobility is different between language areas and between countries. However, in sampling for a number of large telephone speech corpora in the U.S.A. (POLYPHONE; Voice Across America) a special variety called Army Brat was defined, for those speakers who had lived for short periods of time in many different parts of the country. It should be noted that the factor *dialect* does not only affect pronunciation. More often than not, its impact on vocabulary and perhaps also on syntax is at least as important. Of course, the impact on vocabulary etc. can only come to light in corpus collection paradigms which allow the speaker to select his own words. In corpora comprising only read speech this factor should have no effect. Sociolects can be regarded as varieties spoken by a particular social class. A clear distinction between different social classes exists, for instance, in India, where each member of the society belongs to a specific caste.

However, in most cultures it is very difficult to distinguish between social classes. The division into three categories *lower-class, middle-class, upper-class* seems to be most widely accepted for Western cultures. Elaborate schemes have been designed to determine a person's social class using factors such as education, profession, and income. In addition to social class membership a person's sociolect is, of course, also influenced by the linguistic background of the parents and the dialect regions in which he grew up. As is the case for dialects, sociolects may influence not only pronunciation, but also syntax and vocabulary. It is recommended that sociolects should be properly accounted for when collecting new corpora. It has been found that the impact of sociolects on speech behaviour strongly interacts with speaking style. Thus, the speech of a pipe-fitter who speaks in a formal way, may resemble the speech of a salesman who speaks in a casual way (cf. Labov 1972). This phenomenon probably also applies to regional dialects. Occupation-oriented varieties are often termed *registers*.

There is considerable uncertainty on how to treat dialects, and sociolects in corpora collected for developing speech technology, e.g. for developing connected speech recognition systems for use in telephone information systems.

There may be large differences between countries and cultures in what is most appropriate in this respect. Of course, each operational recognition system should be able to handle the range of dialectal and influences present in the speech of upper and middle class speakers produced in somewhat formal situations. The extent to which dialectal influences occurring in less formal speech, or in formal speech of lower class speakers must also be covered will depend very much on the application for which the recogniser is being developed. Another extremely important factor is the social acceptability of strongly dialectal speech in a given situation. Acceptance is likely to differ strongly between regions in a given country.

If telephone applications are designed in such a way that all calls originating from a specific part of the country are handled in a local centre, one may envisage recognition systems which are adapted to the local dialect, provided that suitable training corpora can be collected. When collecting speech corpora over the telephone by soliciting input from randomly selected subjects one should specify strict guidelines for deciding whether or not a specific speaker deviates too much from the "standard" language for him to be included in the corpus. The speech of non-native speakers can be regarded as a special "sociolect". Some non-native speakers may speak the standard language of the country they reside in with only a slight accent, whereas others may speak the standard language with a very marked accent or a poor control over vocabulary and grammar. There seems to be no reason to exclude the former group of non-native speakers from a common speech corpus, whereas the latter group of non-native speakers would preferably be excluded, unless the research is specifically aimed at non-native speech or one wants to build an "all-purpose" speech corpus.

3.6 Final comments

The most important recommendation that must be made is to keep in mind that the purpose of a speech corpus determines the type and design of the corpus. Given a specific aim, one first has to be sure what type of information must be present in the corpus to be collected. Issues to be considered are rather general ones, like the type of recordings (e.g. hi-fi or telephone recordings), but also more detailed ones, like the necessity to collect speech at different hours during the day. In this chapter the type of speech material that can be collected was discussed, running from isolated sounds to spontaneous conversations. In addition, we presented the issues to be considered when the number and type of speakers are concerned.

4 SL corpus collection

4.1 Introduction

The previous chapter discussed the design of spoken language corpus collection. This chapter concentrates on the practical aspects of collecting spoken language material. In the first part, the dimensions of data collection are described to result in a framework for the classification and description of spoken language data collections.

The procedures section contains recommendations for the actual collection of speech data. These recommendations should enable anyone interested in speech recordings to establish a suitable recording environment that will deliver data in a controlled procedure and in an acceptable technical quality. Note that speech data collections always contain "errors": mispronunciations, ungrammatical sentences, new words, technical errors, and so on. These "errors" must be marked, but not removed from the corpus because they contain valuable information. For the development of applications, such errors are required to test the performance and limitations of an application. In speech science, errors are of interest in their own right. The procedures and recommendations in this chapter do not lead to error-free data collections. Instead, they define standards for many aspects of speech data collection which may then be used to explain spoken language phenomena, including the errors.

Clearly, the main object of interest in any spoken language data collection is the speech signal itself. However, additional information can be gathered apart from the basic acoustic speech signal. Whatever choices of speakers, speech material, and recording conditions are made, it is always of crucial importance that the collecting procedure is documented as elaborately as possible. It is good practice to record all possible details about, for instance, sex and age of speakers, type of speech material (isolated words, sentences, discourse, etc.), place of recording (in a laboratory, on location, etc.), type of microphone and recording medium. Although you may not be interested in specific information at the time, it can turn out to be important at a later stage. In that case it is often difficult or impossible to recollect the information you need. And in the second place, a well documented speech corpus may also be used for other research. The following list summarises the most common information sources that may be useful for a speech corpus:

- Examples: acoustic speech signal, laryngograph signal, X-ray data. • Analysis results
- Examples: FFT data, LPC data, filter bank data, pitch extraction, formant extraction.

[•] Transduced signals

- Descriptors Examples: Characteristics of the speakers, or the recording conditions.
 Markers
- Examples: Markers to indicate pitch periods, or the beginning of vowels. • Annotations/Labels
- Examples: Orthographic, phonemic, or phonetic transcriptions. • Assessment parameters
- Examples: Test material, assessment results.

All these information sources must be stored in such a way that potential users of the speech corpus can get access to the speech and the speechrelated data in an efficient and easy-to-use manner.

4.2 Data collection dimensions

Data collection can be represented in the following dimensions:

VISIBILITY: open – secret ENVIRONMENT: studio – on location CONTROL/INTERACTION: random recording – read speech MONITORING: on-line – off-line DATA: single channel – multi-channel

These dimensions are not independent. For example, a secret recording will not be possible in a studio environment, nor will it be possible to do electro-magnetic articulography measurements outside a specially equipped studio.

4.2.1 Visibility: open vs. secret

One concern is that people who are aware that their speech is being recorded change their speech behaviour. Another concern is that speech behaviour is strongly influenced by the social situation in which the speaker performs. Speech produced in a court room, for instance, tends to be more formal in terms of vocabulary, syntax and phonetics than speech produced in one's living room. Also, read speech, especially when the linguistic material consists of isolated words, invokes a more formal and standard pronunciation than free conversation (cf. Labov 1972).

4.2.2 Clandestine microphone recordings

To limit the influence of the fact that the speaker is aware that his speech is being recorded one can hide the microphone and the recorder. Provided that nothing else in the environment alerts the speaker of something unusual, such clandestine or surreptitious recordings, sometimes also called *candid microphone* speech, should be maximally natural. The question arises whether the possible advantage of slightly better naturalness outweighs the disadvantages of hidden recordings, the most important of which probably are the following two very different types: the ethical question of whether clandestine recording is permissible, and the risk of loss of recording quality because on-line monitoring of recording level is extremely difficult. Another possible drawback is that much effort may be spent in recording speakers who, after debriefing, refuse to give consent for the use of the recordings, even for purely scientific research.

A substantial part of the British National Corpus (BNC) has been recorded in situations where at least some persons who participated in a conversation were not aware of the fact that they were being recorded (cf. Crowdy 1993). As far as we know, only a very small number of persons demanded that the recordings in which they participated were erased. It is not known whether the material in the BNC has been analysed to investigate possible differences in recording quality and vocabulary between secret and open recordings.

One obvious way to make sure that the speaker cannot know that he is being recorded is to tape telephone speech. But here again, publishing these recordings as part of a corpus requires that the speaker must be debriefed and asked for consent to keep the recordings.

4.2.2.1 Recommendations

- 1. Use hidden recordings only when there is no alternative.
- 2. For acceptable technical quality, carefully choose the recording environment: tolerable noise, natural setting, and an opportunity to hide the recording equipment.

4.2.3 Visible recordings

In visible recordings, a speaker knows that he is being recorded. This knowledge has an impact on the speech produced: speaker stress phenomena such as a more formal style of speech, a high rate of errors, unnatural pronunciation and prosody can be observed. The environment in which the speaker produces speech, and the task the speaker is asked to perform also determine the style of speech to a large extent.

In many real world situations and applications microphones are present – use of the telephone, dictation, speaking in radio or television studios, traffic guidance, e.g. flight control, etc. Speakers in these situations are used to microphones, and the speech they produce is adapted to the task at hand. In these cases, visible recordings should pose no problems.

In general, if speakers are asked to perform a task that requires some cognitive activity, they quickly forget that they are being recorded. Thus visible recordings can be used even for recording very natural and spontaneous speech if the speaker feels comfortable in the recording situation and is occupied by the task to be solved.

Visible recordings are usually made in recording studios or in environments set up specifically for the recording, e.g. sample office rooms, application simulations, etc. Hence, most recording parameters such as microphone calibration and positioning, recording equipment, can be chosen for maximal quality.

The major advantage of visible recordings is that the recording itself can be monitored on-line to ensure a high technical quality of the recording. Furthermore, speakers have in general agreed to being recorded prior to the recording, so the risk of producing material that may not be used is greatly reduced.

4.2.3.1 Recommendations

- 1. Get written permission for the recording prior to the recording itself.
- 2. Put speakers at ease with a friendly conversation before the recording session starts, preferably in the room where the recordings take place, so that the speakers can get accustomed to their surrounding.
- 3. Explain to the user the significance of the task and the recording procedure.
- 4. Allow the speaker to get accustomed to the environment and the task by doing some practice recordings.

4.2.4 Environment: Studio vs. on location

4.2.4.1 Recording in a studio

Most of the older speech corpora have been recorded in sound studios. Studio recordings have the drawback that most subjects do not feel at home in that environment, with all possible impacts on their speech behaviour. However, as long as the speech to be elicited consists of lists of words, words embedded in carrier phrases, and the like, the "abnormality" due to the unusual reading tasks may outweigh the contribution of the unusual situation. Studio recordings have the advantage of superior signal-to-noise ratio, thanks to the controlled acoustic environment and – at least as importantly – the possibility to monitor recording levels closely, distance of the speaker to the microphone, use of superior but volatile condenser microphones, for example.

One must be aware, however, that "studio" is not a well defined concept. Not all rooms called studio have good acoustic properties. It is not at all unusual to find rooms which have indeed relatively low ambient sound levels, but at the cost of very long or extremely short reverberation times. If studios are used to record large corpora, room acoustics calibration data should be provided with the speech recordings (see Chapter 8 for details on the calibration procedure).

Many (small) corpora designed for basic speech research have been recorded in rooms which were not designed as an audio studio. This is the case with most simultaneous recordings of speech and EMG signals, which are typically made in the research labs of hospitals. Only rarely have these rooms been prepared to provide acceptable room acoustics.

For many applications high quality speech recordings from one speaker at a time are required. These recordings should be free of background noise, including noises made by the speakers themselves. The following guidelines apply specifically to the common situation that speakers are recorded one at a time in a sound studio.

4.2.4.2 Recommendations

- 1. Give the speaker ample time to get accustomed to the studio. Explain the recording procedure in general terms.
- 2. Start recording sessions with a number of practice items to enable speakers to get going.
- 3. At the end of a long recording session, speakers can become hoarse. This can be prevented by taking a sufficient number of breaks in which the speaker can drink some water, or by splitting the long session up into smaller sessions.
- 4. Reduce background noises produced by speakers (e.g. moving their chair, coughing, tapping with their fingers or their feet, turning text pages, etc.) by
 - making the recording conditions as comfortable as possible,
 - explaining to speakers in advance why background noise could interfere with the speech recordings,
 - presenting the text material as clearly as possible, e.g. on a computer screen instead of paper.

4.2.4.3 Recording on location

Corpora recorded in the field have the advantage that the speaker is acting in an ecologically realistic environment. In most cases the price to be paid for this advantage is a substantial loss in signal-to-noise ratio, either because of high ambient noise levels, or limited possibility of monitoring recording levels, distance to microphone, or both. If ecological reality dictates recordings in the field one should nevertheless plan for conditions which allow an audio engineer to monitor the procedure.

Complete ecological validity may not be feasible. For instance, recording a speech corpus in a running car cannot safely be accomplished if the speaker is in the driver's seat.

Two important classes of recordings on location are recordings

- in an actual application, and
- on the telephone.

Recording speech in actual applications (which are based on speech input) is one obvious way to obtain "realistic" speech data. At least in some countries

it is not legally required to advise the user of the service that his speech is being recorded, as long as the recordings are only used for research purposes internal to the company which runs the service. This procedure is probably most often used in pilot versions of an application, where the number of users is limited, so that one may realistically hope to be able to process all recorded speech. Such recordings are necessary to systematically evaluate the success or failure of the speech input parts of an application by relating the speech recordings to the log of the use of the application.

In any case, continued recordings in an application are an efficient means of collecting large amounts of speech data relevant to a given task. Whenever possible, such recordings should thus be made.

Recording speech on the telephone (preferably digital, i.e. ISDN) is suitable for the gathering of limited amounts of speech material from a large number of speakers (POLYPHONE, Damhuis et al. (1994)).

A possible drawback of telephone recordings is the limited bandwidth of the speech signal, typically between 300 Hz and 3000 Hz, which may pose problems for some kinds of basic speech research. For example, the absence of low-frequency components prevents a proper pitch analysis of the recorded speech using methods which rely on the presence of the fundamental frequency (frequency-domain methods, such as the "harmonic sieve", may yield satisfactory results). The absence of high-frequency components prevents for instance the proper spectral analysis of consonants, especially fricatives. Apart from the limited bandwidth of the speech signal, telephone channels can also give a substantial loss in signal-to-noise ratio, especially in the non-western countries where digital telephone systems are not yet commonplace. Even in modern digital telephone networks signal-to-noise ratio suffers from the limited dynamic range that can be accommodated with 8 bit A-law (or μ -law) coded samples.

If these drawbacks are not important for the research goal, recording of telephone speech appears to be a simple way to collect a large amount of speech data in a very short time. For some applications, such as the training and testing of a telephone speech recogniser, a speech corpus with telephone recordings is of course indispensable. It should be emphasised that telephone speech is suitable for many linguistic research projects, including research into most aspects of dialects and regional language variants as well as all aspects related to spoken language syntax and vocabulary.

4.2.4.4 Recommendations

- 1. Recordings in an application are well suited for the collection of large amounts of realistic speech data.
- 2. For each recording in an application, log all additional information, especially system reactions, e.g. success and failure, to provide a basis for performance and fault analysis.
- 3. For telephone recordings, use digital equipment (ISDN).

4.2.5 Interaction and control

In any speech recording, some control over the recording process is needed. The type and amount of control can be characterised by the type of recording:

- random recording
- spontaneous dialogue recording
- interview recording
- read speech recording

The degree of interaction and control clearly is determined by the communication situation of the recording. In a face to face communication situation, the speaker can be instructed directly. On the telephone, interaction and control is restricted to oral communication.

4.2.5.1 Random recordings

In random recordings, all or some of the parameters of a recording, namely the recording time, its duration, the type of speech recorded, etc. are determined in an arbitrary and randomised procedure. In general, the speaker does not know that he is being recorded, and neither does the experimenter know who or what is being recorded. Other than the original setup of the experiment, there is no control over the recording, and no interaction between the experimenter and the speaker.

Such random recordings typcially are: "record speaker X from the speaker population for Y minutes every Z hours" (as in the BNC), or "record the news on channel X every Y minutes", or "record the speech of microphone number X in the airport tower control room". In the first example, X, Y, and Z are appropriate random numbers, in the third example the speaker using this particular microphone is randomly selected. In all three examples, there is no control over the contents of the speech that is being recorded. Examples of random recordings are the ATIS recordings of air traffic control and Switchboard of telephone services.

Random recordings are used a) to gather huge amounts of task-oriented speech, b) to reduce the subjective bias of the experimenter concerning the selection of speech to be recorded, and c) to eliminate speaker stress that results from the knowledge of being recorded.

4.2.5.2 Spontaneous dialogue recordings

In spontaneous dialogue recordings, once the recording equipment is set up, the speakers have been briefed, and the dialogue has begun there is no more control exercised by the experimenter (except perhaps to terminate a dialogue). In a spontaneous dialogue, the participating speakers mutually control themselves by taking turns (exclusively or overlapping).

The experimenter may or not be a speaker in the dialogue. The dialogue may be focussed on a given topic or restricted to a maximum duration, or it may be completely unrestricted.

In the VERBMOBIL project, negotiation dialogues are recorded. The task to be solved is to find a date for a business trip for two partners. The dialogues are recorded in a studio with two separate rooms with visual contact through a window, and turn taking is controlled via a button (Hess et al. 1995).

Spontaneous dialogue recordings are used a) to collect speech for the analysis of dialogue structures, b) to obtain natural speech with the full set of prosodic phenomena, and c) in role-play where task-specific speech is simulated.

4.2.5.3 Interview recordings

In an interview an interviewer prompts a speaker to produce speech and leads the speaker through the interview. The interviewer can be the experimenter, a trained human interviewer, or a speech computer.

The interaction between interviewer and speaker begins with a briefing before the interview. During the interview, the interviewer prompts the speaker to respond to questions, to repeat words or sentences, or to discuss a given topic.

The interviewer has various control instruments at his disposition: he can interrupt the speaker, ask for repetition, change the order of topics in an interview, skip or insert new topics, and deal with topics in various degrees of detail.

The influence of the interviewer on the course and the resulting speech of an interview is strong. A good interviewer can establish a relation with a speaker in such a way that the original goals of the recording can be achieved. In many cases, a few interviews by a trained interviewer can produce the same amount and quality of speech material as very large random or spontaneous speech recordings. Furthermore, because of the strong interaction control by the interviewer, speech recordings are focused. Finally, the technical quality of the speech recording is much easier to control in an interview than in random or spontaneous dialogue recordings.

Computers can also be used as interviewers. The advantages here are that the influence of the interviewer on the speakers is the same for all speakers, allowing the direct comparison of interview recordings, and that multiple measurements can be made during the recording already, e.g. timing measurements. The disadvantage is that the computer can only follow a predetermined script and is thus not able to adapt to situations not foreseen during the design of the speech collection or the script. Furthermore, many speakers feel uncomfortable when they know they are talking to a machine. Interview recordings are used a) to elicitate speech in a rather controlled way, not necessarily leading to unnatural speech, b) to collect speech via the telephone, and c) to perform fully automated speech collections of large speaker populations.

4.2.5.4 Read speech recordings

In read speech recordings, the speaker is asked to read exactly what is presented to him or her. The text to be read may either be printed on paper or presented on a computer screen. In the first case speakers tend to change their articulation according to the perceived structure of the text, e.g. lower the voice at the end of an enumerated list. Presenting text on a computer screen avoids this problem, but many speakers are intimidated when facing a computer screen.

In read speech recordings, the degree of control is very high. The text to be read can be generated according to predetermined criteria, e.g. distribution of phonemes, vocabulary, etc. During the recording, each utterance of the speaker can be checked directly for errors, and, if an error is found, the speaker is asked to re-read the text.

Read speech is not spontaneous speech. Nevertheless it is close to natural speech in some specific speech styles, e.g. dictation.

Read speech recordings are used a) to guarantee that the speech material has a certain content, b) to record the phenomena specific to read speech, and c) to monitor the recording very closely, e.g. in multi-channel recordings.

4.2.5.5 Speaker prompting

Speaker prompting is used to elicit directly from a speaker a certain type of speech data, e.g. numbers, dates, times, etc. Such data is much more difficult to obtain in dialogues or role-play. The major problems with speaker prompting are a decrease in the spontaneous quality of the speech, ambiguous prompts that lead to unexpected responses, and the rigid structure of a prompting script, i.e. a sequence of prompts. Furthermore, speakers tend to imitate the original prosody when they are asked to repeat an utterance (thus there should be different prompts for the same text so that various prosodic patterns for the same text are recorded).

However, in many applications prompting a user for input is a natural situation, and thus the decrease in spontaneity in the speech is highly welcome. Four types of prompts can be distinguished:

- information or feedback prompts, e.g. You have now completed half the questionnaire.
- instruction prompts, e.g. Please read the number under topic 1).
- questions, e.g. what day of week is today?
- repetition prompts, in which a speaker is asked to repeat what has been prompted, e.g. *three four five.*

The possible responses to prompts may vary greatly. It is thus advisable to instruct the speaker which responses are expected, e.g. "Please answer the following questions with yes or no". However, restricting the set of allowed responses too strongly will lead to unnatural speech.

In face to face communication situations, there is an influence of the interviewer on the speaker even when the catalog of prompts is fixed, e.g. in a prompt sheet. Visual communication, deictic references (like pointing with a finger to an item to be read), play a significant role. The interviewer guides the speaker through the script and may immediately correct any errors.

In telephone recordings prompts may be output by a computer or a human interviewer. The advantage of computer prompting is that all speakers hear identical prompts. One disadvantage is that a computer based prompting system strictly follows a predetermined prompting script and may not notice that the speaker is not responding correctly. Computer prompting scripts should thus not take longer than 15 minutes, and the script should be divided into several small units. Between the units, feedback should be given to the user to inform him of the status of the recording.

Human interviewers immediately realise whether a response from a speaker is correct, and they are able to correct wrong responses immediately. However, each prompt is an individual utterance so that variations among responses may also result from the prompts.

Each prompting script, should be thoroughly tested before the actual recording of data. The test participants should be candidate speakers, and test conditions must be as similar as possible to those of the actual recording. In the case of computer prompting, it is useful to have a prompting simulator which can be adapted to new prompting scripts easily.

- 4.2.5.6 Recommendations
- 1. Formulate prompts so as to avoid ambiguities.
- 2. Restrict the number of allowed responses.
- 3. Test prompting script on 1% (minimum 10 speakers) of the candidate speakers under realistic conditions.
- 4. Give sufficient feedback to the speaker (for instance, after a third, a half, and three quarters of the script).
- 5. Add dummy items at the beginning and at the end of text pages. Also, each speaker should have a different random ordering of test items so that positional effects on the pronunciation can be levelled out.
- 6. Add one or more repetitions of the test items.
- 7. Use a VDU for presenting the prompting material when recording in a studio or on the experimenter's premises. (When a VDU is used, care should be taken to avoid reflections from the screen and it should be checked that the monitor does not produce audible noise of its own).

4.2.6 Monitoring and validation

Monitoring is the task of controlling and modifying technical and phonetic characteristics on-line, i.e. during the course of a recording. Validation relates to an off-line (or *post hoc*) technical or phonetic evaluation of the material recorded.

Monitoring can best be applied in studio recordings of read speech, and, to a limited extent, to interviews or dialogues. Technical characteristics such as the recording level, selection of recording channels, and even adapting a computer based script to a speaker can be varied while the recording is going on. Monitoring the phonetic characteristics of a recording concerns the quality of the speech – pronunciation, speaking style, etc. Two on-line monitoring paradigms can be distinguished: one in which any deviation or error is signalled to the experimenter only, and another one in which also the speaker is informed that a particular error has occurred.

Monitoring is the only practical paradigm that guarantees that the corpus will indeed contain exactly the items and the number of repetitions planned for during corpus design. The procedure has one very important limitation, that should not be underestimated: it will yield a corpus which is (virtually) completely devoid of dysfluencies, out-of-vocabulary words, coughs, sneezes, etc. Cleaned-up corpora of the type implied here have misled engineers to think that speech recognisers had reached performance levels sufficient for actual applications. What they failed to realise - due to the absence of these phenomena from the training materials – was that in real life dysfluencies etc. abound, and that these phenomena may be more important in determining the real life performance of speech recognisers than recognition error rate on a clean corpus. For this reason it is strongly recommended to use post hoc transliteration whenever that is possible. In making this recommendation it is acknowledged that recording dysfluencies etc. makes no sense in recording speech material for carefully designed perception experiments.

Some characteristics of recorded speech can only be evaluated after the recording has taken place. In the technical domain, such characteristics are the signal-to-noise ratio for the whole material, and an analysis of noises that were recorded together with the speech. The phonetic characteristics include an anlysis of the items produced, an orthographic transliteration or phonemic transcription of the speech signal, and segmental and prosodic descriptions.

Post hoc validation was employed in collecting some (very) large corpora like Voice Across America and POLYPHONE. It is also used in the German VERBMOBIL corpus collection, where the dialogue recordings are transliterated orthographically and then transcribed relative to a given citation form (Hess et al. 1995).

4.2.7 Multi-channel recording

For some applications one must record a number of physiological signals besides the acoustic signal, such as a laryngograph signal, an electromyograph signal (EMG), air pressure or flow in the vocal tract, articulatory parameters, X-ray data, etc. The major drawback of recording such additional signals is that speakers have to be bothered with measuring equipment, such as a strap with electrodes round the neck in the case of laryngographic recordings. One should be aware that the measuring equipment may interfere a natural speech production. Therefore, it is recommended to use additional signal recordings only for basic speech research and for specialised purposes, such as examination of voice pathology, and else confine oneself to the basic acoustic signal. For some applications (audio-visual analyses) it may also be useful to make video recordings. The following are examples of ancillary information channels:

- Electroglottography (Laryngography)
- Electropalatography
- Electromagnetic Articulography (EMA)
- Cineradiography
- Air-flow measurements

Most of these multi-channel recordings require a high technical effort. The placement of sensors may disturb the speaker (X-rays are even dangerous), and due to the considerable effort involved, only few corpora of these kinds of measurements exist.

4.2.7.1 Laryngography

In a laryngograph recording a pair of electrodes is attached to the throat of the speaker on each side of the thyroid cartilage (Adam's apple). This sensor produces a signal proportional to the amount of contact between each vocal fold, e.g. during phonation.

Laryngography recordings were taken at the Eurospeech 93 conference in Berlin, where speakers were recorded during their presentations. The data are available in the TED corpus from the Bavarian Archive for Speech Signals (BAS; see Appendix N) and LIMSI.

REQUIREMENTS: Laryngography sensors, DAT tape recorder or computer interface; 8 bit quantisation approximately 10 kHz sample rate. Laryngograph recordings are also included in the EUROM-1 corpus.

4.2.7.2 Electropalatography

Electropalatography registers the contact of the tongue with the hard palate during articulation. The speaker places a customised thin artificial palate

in his mouth. This artificial palate contains an array of electrodes which record contact with the tongue.

The data recorded by each electrode is combined to a two-dimensional representation of the palate at any given point in time.

REQUIREMENTS: Artificial palate individually tailored to a speaker, multichannel recording device, e.g. computer with a suitable interface; 64 bit quantisation (i.e. typically an 8×8 array), sample rate 200 Hz.

4.2.7.3 Electromagnetic articulography

Electromagnetic articulography (EMA) measures the movement of the tongue and other articulators through tiny induction coils attached to the tongue. The head of the speaker is enclosed by a helmet which usually holds two (or more) coils that create an electromagnetic field; The signal induced in the coils on the tongue is proportional to the distance from the transmitter coils on the helmet.

The EMA provides essentially the same kind of data (for parts of the vocal tract only, because coils cannot be placed on the larynx) as the microbeam X-ray (see Chapter 8) but uses a different technology.

REQUIREMENTS: Articulograph, multi-channel recording device, e.g. a computer with a suitable interface; data rate depends on the quantisation, the number of sensors, the number of transmitters (typically 10 sensors and 3 transmitters), and the sample rate (typically 250 Hz).

4.2.7.4 Cineradiography

X-ray measurements are rarely performed today because of the health hazards they impose on the speaker. However, early recordings are still available on film or, in digital format, on laser disk (Bateson at ATR, Japan). X-ray measurements show the modification of the articulatory tract during articulation. The movement of the jaw can be seen clearly; tongue and lip movement are often less clear due to the fact that they do not show up very clearly on X-ray. The movement of the vocal folds is too fast to be recorded at the slow frame rate of film recordings.

REQUIREMENTS: Seldom performed.

4.2.7.5 Air-flow measurements

In air-flow measurements the speaker wears a mask (usually designed to separate oral and nasal airflow). Flow is usually derived from the pressure drop across a wire-mesh located in a flow head mounted in the mask.

The measurements yield data on the speed, direction, and volume of air flow. Depending on the type of sensor and attachment, the measurement requires that the speaker does not move during articulation.

REQUIREMENTS: Air flow sensors, data acquisition hardware. The data rate depends on whether phonatory components of airflow need to be captured.

4.2.7.6 X-ray microbeam

X-ray microbeam provides two-dimensional movement data (usually in the mid-sagittal plane) of selected fleshpoints on the tongue and other articulators. It uses a point-tracking technique to reduce the radiation exposure to the subject to acceptable levels.

REQUIREMENTS: The equipment is only available at the dedicated microbeam facility in Madison, Wisconsin. Data rate: each fleshpoint is tracked at about 100–200Hz. Typically about 10 fleshpoints are tracked simultaneously.

4.2.7.7 Nuclear magnetic resonance imaging

Nuclear magnetic resonance imaging is a static (up to now) imaging technique with very good resolution of the soft tissues in the vocal tract. Slices can be freely chosen i.e sagittal, coronal, etc.

REQUIREMENTS: a friendly hospital; sample rate < 1Hz, image resolution 256 \times 256 pixels (typical, with 8 bits pixel depth)

4.2.7.8 Ultrasound imaging

Ultrasound imaging can be used for obtaining sagittal and coronal images of the tongue(for those locations on the tongue where no air intervenes between transducer and tongue; the transducer is usually held under, and moves with the jaw).

REQUIREMENTS: Ultrasound machine. The data is usually stored as standard video data. A frame-grabber is needed if data is to be digitised.

4.3 Procedures

The following sections describe procedures which have been used in practical speech recordings.

4.3.1 Equipment

The basic equipment needed for speech recordings consists of

- microphone,
- amplifier and/or processor,
- recording device,
- storage devices.

The choice of equipment depends on the choices along the dimensions visibility, environment, communication mode, and on the data to be recorded. The quality of the recording channel itself (microphone and recording medium) is determined by three characteristics: signal-to-noise ratio, bandwidth, and dynamic range.

For every speech recording a *log* or *journal* should be kept. It contains the essential administrative information about recording setup, personnel involved, speaker data, and recording time and date.

It is necessary to store at least the following data for a recording session:

- recording date and time,
- recording engineer,
- equipment used,
- ${\rm speakers},$
- recording environment.

Recording time and date and the recording engineer are independent of the number of speakers or channels recorded. The environment, speakers, and equipment may differ for each channel and thus should be written down separately for each channel. A separation of recording dependent and channel dependent data is thus advisable, and this separation should be made explicit in the layout of a form or a database structure (see Table 4.1).

rec_id	date	time	engineer	remarks		
M0127D	22.01.95	17:10	CSC	VM scenario "Time Table"		
				April–June calendar		
id	channel	type	recorder	micro-	environ-	speaker
				phone	ment	
342	left	audio	sony DAT	HMD416	studio	UAA
343	right	audio	sony DAT	HMD416	office	AJB
344	Ā	electro-	c:/epg/300	palate UAA	studio	UAA
		palato-				
		graphy				

Table 4.1: Possible layout for recording session information

4.3.1.1 Microphone

Microphones can be classified by their

- directional properties,
- transducer principle, and
- position relative to the speaker.

A well-documented speech corpus should contain data about the microphone, such as make and type (condenser, dynamic, etc.), position of the microphone relative to the speaker or speaker's mouth, possible calibration procedures, etc. (see Chapter 8 for further information).

Unidirectional microphones are sensitive to the direction where the sound comes from. Unidirectional microphones are preferred when a single speaker is recorded in a laboratory environment.

Omnidirectional microphones are not sensitive to the direction where the sound comes from. They are suitable for recordings on location, or when the speaker is moving, e.g. working, walking, or driving a car. Omnidirectional microphones may be used to record several speakers if it is guaranteed that their turns will not overlap.

Microphone arrays are a possible means of sound source location. Computer controlled microphone arrays can focus on a single sound source, therewith improving the effective signal-to-noise ratio dramatically (cf. Flanagan et al. 1991).

The *electrical transducer principle* is a second dimension along which microphones can be distinguished. For most purposes in speech research the differences between the tranducer principles is not very important, with the exception of carbon button microphones. Carbon button microphones are used in older telephone handsets. They may distort the frequency response of the signal quite considerably. Moreover, their transmission properties may change significantly over time. Electret microphones are more stable than carbon button microphones. They can have almost flat frequency response in the telephone bandwidth, at least in principle. The actual frequency response of a microphone depends primarily on the acoustic properties of the case in which it is encapsulated. Badly designed handsets therefore can have bad frequency response characteristics, regardless of the use of electret transducers. For basic research into the characteristics of the glottal sound source the phase distortion of the microphone is as important as its amplitude response. For virtually all other research and development purposes in speech phase response is immaterial.

Finally, the *microphone position* relative to the speaker's mouth can be used to distinguish types of microphones.

Headset microphones usually are attached to headphones via an arm. The position of the microphone relative to the articulatory tract is fixed, and the speaker is free to move the head. However, the microphone has to be

positioned very carefully to avoid noise through breathing, and speakers often feel uncomfortable with a headset. If the task to be solved by the speaker is sufficiently complex, unconsciously produced gestures such as lip smacking, scratching one's head, rubbing one's chin, etc. may produce significant noise, especially if the headset is touched.

Close-up microphones are attached to the speaker's clothes, usually on the chest. The microphone does not disturb the speaker and it is quite close to the articulatory tract. However, the distance of the microphone varies greatly with body movements, and new noise sources, e.g. rustling of clothes, are introduced.

Table-top microphones usually are unidirectional microphones placed approximately 50 cm away from a speaker. The microphone does not disturb the speaker, and the distance of the microphone varies only very little with body movements. However, with more than one speaker in a room there is little channel separation, and new noise sources, e.g. interference from room echo, tapping on table, movements of prompt sheets, are introduced.

Room microphones are omnidirectional microphones that are placed in specified positions in a room. They are independent of speaker position and can be hidden completely. However, there is little (if any) channel separation, and surrounding noise interferes with the speech signal.

4.3.1.2 Recommendations

- 1. If acceptable in the recording environment, and for optimal acoustical quality, use headset microphones.
- 2. Place the microphone slightly to the left or the right of the mouth and a bit below the lower lip to avoid breathing noises. Take care that no cables touch the microphone arm, and that the speaker is comfortable with the headset.
- 3. With headsets, have the speakers control their hands, e.g. by pressing a button or holding a computer mouse so that they do not touch the headset.
- 4. Take care that the attached cable does not tap against any hard surface. The sound is transmitted to the headset.

4.3.1.3 Amplifier/Processor

The signal coming from a microphone must be amplified to be recorded. In many cases, some processing is also needed, e.g. analog to digital conversion, transformations for different encoding schemes, filtering to reduce noise, etc. Some processing steps have to be performed only once, e.g. analog to digital conversion. Others will be performed repeatedly, e.g. the transformations for different encoding schemes.

4.3.1.4 Recommendations

- 1. Define a standard setup and procedure for all the steps from recording the signal to storing it.
- 2. Choose *de-facto* accepted standards for the storage formats, and use standardised conversion tools. For normal speech recordings the standard quantisation and sample rates should be 16 bit (linear encoding) and 16 kHz, and 16 bit 8 kHz for analog telephone speech. For ISDN telephone recordings, one should use the ISDN standard 8 bit A-law encoding at 8 kHz sample rate.
- 3. Use the same equipment wherever possible (and appropriate).

4.3.1.5 Recording device

Basically, there exist two types of recording devices: tape drives, and computers with hard disks. Recordings to tape are either analog (audio tapes, compact cassette, video tapes) or digital (DAT), whereas recordings to hard disk are always digital.

Ongoing development in the field of audio technology has shifted the emphasis away from analogue recording media to digital recording media. The traditional recording medium has been the reel-to-reel magnetic tape. Apart from a relatively poor signal-to-noise ratio of typically 60–70 dB, this medium suffers from mechanical problems such as flutter and wow. Moreover, the quality of an analogue speech recording severely degrades after it has been copied repeatedly. Because of these drawbacks, it is strongly recommended to use digital media for the recording of speech. The most widespread digital medium for recording of speech signals is the DAT (Digital Audio Tape). This medium is strongly recommended. Recordings are made on two channels with standard sampling frequency of 48 kHz, and 16-bit resolution. Another option, that can only be used in a laboratory environment, is to record the speech directly on a high capacity computer disk. Two other digital audio media, the CD-ROM and WORM (Write Once Read Many), are less suitable for speech recording, because they cannot be erased. That is, data (for instance, speech recordings) can only once be written to a CD-ROM or WORM; afterwards, the stored data can be read as many times as one likes (compare a grammophone disc). The CD-ROM and WORM are especially useful for the permanent storage of selected recordings in a database.

The recording devices can be characterised according to the following criteria:

- portability,
- capacity, and
- ease of use.

The *portability* of a recording device is determined primarily by its size and weight, and secondarily by its operating requirements, e.g. power supply, environmental conditions, etc.

Tape drives, analog or digital, come in all sizes, including Walkman-sized DAT recorders. Usually, tape drives are optimised to record or playback signals, i.e. they do not produce very much noise themselves. Portable tape drives usually have only a reduced set of features, they can operate on batteries and are quite immune to adverse environments (some are even water resistant). Non-portable tape drives offer more features (e.g. remote control, manual setting of recording parameters, computer interfaces), require a permanent power supply and operate in the usual office environments.

Computers too can be divided into portable and desktop computers. In general, they produce significant noise during operation (hard disk spinning, keyboard clicks, system alerts etc.) and must thus be shielded from the signal to be recorded. Furthermore, sound cards in computers are subject to interference from other devices inside the computer, e.g. noise from the bus, the processor etc. Portable computers are about the size of an A4 book and weigh approx. 2 kg. At present, only high-end portable computers are equipped with the signal processing facilities (e.g. signal processor, 16 bit quantisation, sample rate > 8 kHz) required for speech recordings.

The capacity of tape drives is almost unlimited because full tapes can be replaced by empty ones quickly and at low cost. Typically, an analog compact cassette holds about 90 minutes of stereo signals, a video cassette up to four hours, a digital DAT tape up to two hours.

The capacity of computers for speech recording is mainly limited by the capacity of the hard disk. A 1-Gigabyte disk can store approx. 8 hours of mono signals (16 bit quantisation, 16 kHz sample rate). Such disks are becoming common on many desktop computers and even in portable computers, so that hard disks are suitable recording devices for very many speech recordings already. The major limitation of recording to hard disks is that the hard disk cannot simply be exchanged against another one. This means that the data on a hard disk has to be saved to some backup medium, e.g. magnetic tape or CD-ROM.

Ease of use must be seen under two aspects: first, the ease with which the device can be used to perform the recording; second, the ease with which the recorded data can be accessed for further processing.

Tape drives are easy to set up and speakers are used to them. However, especially for analog recordings, it is quite cumbersome to access recordings for further processing. The appropriate tapes have to be located and the tape drive has to be attached to a computer.

Computers as recording devices are still uncommon. They require the expertise of an engineer to be set up correctly, and speakers are easily distracted by the presence of a computer. However, computers offer significant advan-

tages over tapes: recordings can be fully automated, administrative data is collected together with the recordings, and data is available immediately, either for control purposes or further processing.

4.3.1.6 Recommendations

- 1. Use digital recording devices.
- 2. Use a computer for the recording to automate recording procedures and for easy access to data for further processing.

4.3.2 Management

The management of a speech data collection involves

- recruitment of speakers,
- scheduling of recordings and speakers,
- cost control.

Speech data collections are very expensive, both in terms of money and resources. It is thus recommended to collect, besides the speech signal itself, as much administrative data as possible, and to design procedures that can be reused in other data collections.

4.3.2.1 Speaker database

Speakers should be thought of as a primary and very valuable resource in speech recordings. It is therefore advisable to build a *speaker database* which contains for each speaker

- a unique speaker id,
- administration data (name, address, telephone),
- personal information (place and date of birth, languages, education, etc.),
- physiological data (sex, size, weight, etc.),
- speaker history (list of recordings, etc.),
- remarks.

Preferably, such a database is implemented using a database management system on a computer (see Appendix H for details on DBMSs).

This way, data can be entered easily during the preparation of a speech recording. Such a database can also be held on forms in a folder, but then the extraction of speakers according to specific criteria other than the primary ordering criterium is difficult and error-prone.

4.3.2.2 Speaker recruitment

The recruitment of speakers should have two goals: provide a sufficient number of speakers for a given speech data collection, *and* provide sufficient information about the speakers which can be used to build or extend a speaker database.

Speaker recruitment can be characterised along the following dimensions:

- number of speakers,
- speaker qualifications,
- demographic distribution,
- speaker availability (once/repeatedly; within short/long period).

Recruiting a small (i.e. 1 to 5) or medium (5 to 50) number of speakers is no problem. Depending on the requirements, colleagues, friends, and relatives can be asked to participate. However, one cannot expect any demographic balance in small sets of five or less speakers. The advantage of using friends and relatives is that they may be available for a long period of time, and that they in general can be used for more than one recording.

The recruitment of a large to very large number of speakers is completely different from that of a small number of speakers. Accessing the speakers, scheduling their recordings, evaluating the recordings, and storing the data become such large tasks, that they cannot easily be performed by a single person. Accessing a large number of speakers requires either

- a contact address,
- a public call, or
- a hierarchical recruitment scheme.

Contact addresses, i.e. telephone number or postal address, are expensive: address brokers charge for each address bought, with the risk of the address being useless (the address is wrong, the person is not willing to cooperate, etc.) entirely upon the buyer. Market research institutes have large address databases from which they can select subsets according to specific criteria, but they in general do not give away these addresses. Although addresses allow persons to be contacted directly, e.g. through mail, telephone, or interviewer visits, the rate of return is rather low, typically in the range of less than 5% for mail, 25% for telephone, and 50% for interviewer visits. Public calls for participation, e.g. newspaper advertisement or article, Internet posting, radio or TV announcement, may reach a very large audience. In many cases, a public call can be arranged at little expense – newspapers, especially the science editors, are willing to cooperate, Internet postings are virtually free, and radio or TV announcements are affordable). The rate of return is usually very low (less than 1%) but this is compensated for by the sheer size of the audience reached. However, the means to determine the response to a call for participation are limited. Also, the number of callers will not be evenly distributed over time (most people will call immediately after having received the call), which may cause capacity problems. People responding to a public call for participation are highly motivated; however,

this does not hold for the population as a whole and thus introduces a bias. In *hierarchical recruitment* the task of recruiting m speakers is divided into n tasks of recruiting m/n speakers. Hierarchical recruitment works well if the

burden of recruiting speakers can be mapped to some real-world hierarchy, e.g. the employee hierarchy in a company. The rate of return strongly depends on the success of a person persuading others to participate.

In all three recruitment strategies, incentives may help to increase the motivation to participate and thus the rate of return. Incentives can either be gifts (e.g. telephone cards) or the participation in a lottery with a grand prize. However, such incentives clearly make the recruitment of speakers even more expensive.

4.3.2.3 Recommendations

- 1. Speaker addresses are a valuable resource. One goal of the speaker recruitment should thus be to establish or extend the database of speakers.
- 2. In a project with a large project partner, e.g. a large company, the hierarchical recruitment scheme should be tried.
- 3. If sufficient time is available, then a mix of a public call for participation and contact addresses is promising: a public announcement provides an address from which further information, prompt sheets, etc. can be obtained. Any such request is then registered and the person can be addressed directly.
- 4. If time is short and manpower is limited, then speaker recruitment via a market research institute is advisable. This is especially true if there is a high degree of interaction between the speaker and the recording institution, e.g. agreements to be signed, posting of incentives, etc.

4.3.2.4 Scheduling speakers

Scheduling speakers is important to make optimal use of the recording capacities within a given period of time. Proper scheduling avoids speaker frustration (caused by having to wait, ever-busy telephone lines, etc.) and allows a maximum number of recordings within the given recording capacity. If speakers are recorded in a studio, a time slot is reserved for each speaker. This time slot must be sufficient for

- the technical setup of the recording,
- the registration of the speaker,
- the instruction and accustomisation of the speaker,
- the speech recording session itself,
- a cleanup procedure (backup of data, reset of technical equipment, etc.).

In general, five minutes for each of the side-tasks should be sufficient. If speakers have to travel far then it is almost inevitable that some of them come late or do not appear at all. In such cases it is advisable to have some speakers available upon short notice. In any case there must be a person responsible for the scheduling, and this person must be reachable directly by telephone. For telephone recordings, the number of speakers calling at any one time must be matched to the capacity of the telephone equipment. If potential speakers do not get through because of busy lines, they are likely not to retry. Furthermore, telephone recordings should be possible 24 hours a day, or it must be clear to callers that the service is operational only for a specific period during the day. Note that recording 24 hours a day requires that the recordings be performed automatically because only in rare cases will human operators be available for 24 hours. Again, speakers must be able to reach an operator via telephone, e.g. to report problems or make suggestions.

4.3.2.5 Recommendations

- 1. Spread the number of speakers evenly over the recording period, but leave the last 10–15% of the period as a reserve to cope with delays and interruptions.
- 2. In studio recordings, assign a time slot to every speaker. Once a speaker is in the studio, record as much data particular to a speaker as possible (name, address, availability for future recordings, physiological data, health history, etc.).
- 3. In telephone recordings, ensure that the recording system is operational without supervision, and that the expected number of speakers is well within the capacity of the telephone system (occasional bursts of activity must not disrupt the recording system).
- 4. One person should be responsible for the scheduling of recordings. This person must be reachable directly via telephone and fax.

4.3.2.6 Cost

The cost of a speech recording is determined by the cost for personnel and equipment and by the period of time. The total cost estimate is defined in a budget, and at given times the actual expenditures are compared to the budget plan.

A speech recording project is usually defined by scientifically trained experts, e.g. speech engineers, phoneticians, etc. Only rarely are there people with expert finance and budget knowledge in a project. Hence budgets often are rather broad estimates, and many hidden costs are easily overseen.

The minimum personnel requirements for speech recordings (of a large number of speakers) are a project administrator and supervisor, and a system operator; both should be available for the whole recording period. Depending on the speech recording setup and the processing of the signal data, interviewers, scientific personnel, and temporary collaborators are necessary.

The administrator is responsible for the budget and the supervision of the project as a whole, the recruitment of speakers, scheduling of recordings,

and the organisation of the data evaluation. The system operator is responsible for the technical and data processing aspects of the speech recording, i.e. the setup of equipment, storage and backup of data, etc. Interviewers are needed for speech recordings in face to face communication situations. A first evaluation of the technical quality of recordings can be performed by rather unskilled personnel, whereas further processing, e.g. the transliteration or a phonetic segmentation and labelling of utterances require trained experts or scientific personnel.

The cost of personnel is the sum of salary and related infrastructure (room, desk, computer, telephone) and working materials costs. In many cases existing resources can be reused, but it should be clear that they have to be accounted for in the budget.

The cost for equipment consists of the acquisition and maintenance costs. Again, in many cases existing equipment can be reused and it must be accounted for in the budget. Maintenance costs are significant cost factors which often exceed the original acquisition costs. For time-critical projects, maintenance contracts with a guaranteed repair time should be considered. A speech recording can be divided into the following phases:

- initialisation and test phase,
- $\bullet \ \ {\rm preparation},$
- recording,
- evaluation,
- clean up.

All phases are strictly sequential except for recording and evaluation which can be executed in parallel.

Initialisation and test, preparation and cleanup take roughly constant time. The initialisation and test phase must be considered very important because wrong decisions here will affect the rest of the project. Preparation can be short if the initialisation and test phase results in a good procedural setup. The duration of the recording and evaluation phases depend directly on the number of recordings. As a rough estimate, double the speaking time (prompts and responses) to get an estimate of the time needed to perform an individual recording (speaker instruction, cleanup, etc.). Depending on the quality of the evaluation, the time needed for evaluation may be double (technical evaluation) to ten times (phonetic evaluation and transliteration) the speaking time.

4.3.2.7 Recommendations

- 1. Define a detailed budget for the project as a whole and the speech recording tasks.
- 2. Assign a project supervisor and a system administrator; both persons should be available throughout the project.

- 3. Exploit parallelisation of tasks.
- 4. Include overhead in your time calculations: for recordings, 100% overhead on speaking time, for evaluation 100–1000%.

4.3.3 Wizard of Oz

In a Wizard of Oz experiment, a speaker believes that he is communicating with a computer, while in reality the computer is simulated by a human. Wizard of Oz experiments are particularly useful in the simulation of computer-based spoken language systems for a clearly limited task, e.g. information retrieval systems like telephone information services, train enquiries, automatic translation of spoken language, etc.

The setup of a Wizard of Oz experiment must be convincing. The speaker must neither over- nor underestimate the experiment. This requires that the speaker believes that a computer could potentially handle the task, and that the task is significant enough to require the use of a computer. Modifying the speech produced by the human impostor in such a way that it sounds machine-like, e.g. monotonous, mechanical, etc., may make the experiment more convincing. If the generation of a response to the speaker takes some time, give the speaker a convincing explanation of the delay to be expected; a progress indicator (which may be fake) on the computer screen will keep the speaker interested and waiting for a response.

A Wizard of Oz experiment should be developed in at least two phases: in the exploratory phase the human impostor is free (within the limits of the task) in the responses to the speaker, and all responses are logged. For the production phase, the log files are evaluated to result in an abstract script which covers the allowed and desired subtasks. The human impostor must then adhere to this abstract script for the experiment.

For a more thorough treatment of Wizard of Oz experiments, see Chapter 13.

4.3.3.1 Recommendations

- 1. Carefully hide the human "impostor" from the speaker.
- 2. Record both the speaker and the human impostor, and log all responses of the computer.
- 3. Clearly explain to the speaker the limitations of the task and the computer.
- 4. Develop a Wizard of Oz experiment in phases: in an exploratory phase get the experiment to work, in a production phase do the actual recordings.

4.3.4 Legal aspects

There are differences between countries in legal objections against collecting, storing and disseminating demographic and personal characteristics of speakers in a corpus. In all countries name and address data may freely be stored and published, as long as it is guaranteed that no other data about

the person can be linked to name/address. Limitations in this respect must be carefully checked with legal advisers in the country or countries of interest.

Another legal issue that must be considered is the consent to record speech and subsequently to make the recordings available to other parties. In legal terms, making available recordings is probably equivalent to publishing them. There are likely to be differences between countries in legislation about recording speech. In some countries recording is legal as long as one of the parties involved in a dialogue gives explicit consent. The law may only state that recording for one's own use is legal; publication or any other way of dissemination may be illegal, or bound by much stricter regulation. To avoid problems in this area, it is recommended to have speakers sign a statement to the effect that they know that the recordings are made for later dissemination and publication. If that is not possible, e.g. when collecting large corpora of telephone speech, it is recommended that each speaker is explicitly advised to abort the call if he has any objections against recording and subsequent publishing of the speech to be produced. It is very important to consult with legal advisers about the correct formulation of these statements and advices.

Special care must be taken when recording and publishing corpora with pathological speech or with speech of very young children who are not yet able to give conscious consent for publishing. Especially with rare pathologies it may be very easy to trace the speech back to the patient, even if name and address data are not coupled to the data about the pathology. With pathological speech of mentally healthy adult patients a carefully formulated written and signed consent form may be sufficient. With speech of very young children carefully formulated consent forms signed by the parents are necessary.

Legal issues (and ethical ones as well) are especially relevant in the case of surreptitious recording of speech. In order to circumvent the observer's paradox (see Section 3.4.1 under "Spontaneous speech"), researchers might want to resort to surreptitious recording of speech. This could, for instance, be done with a concealed microphone "in the field" or by tapping telephone conversations. Invading the privacy of someone's personal speech might be regarded as illegal by the authorities. And even if it would not be illegal, it might still be regarded as unethical by many people. As far as we know, no linguist has ever been tried for recording speech data surreptitiously. However, to minimise the risk of breaking the law in any way, and to conform to ethical norms as much as possible, the following guidelines can be taken into consideration:

• Be sure that recordings are made only to gather speech data that could not be otherwise obtained. Also remember that any possible tortious or criminal motive would certainly render a surreptitious recording illegal.

- To protect the privacy of persons who were surreptitiously recorded as much as possible, do not disclose their identities to anyone, do not use real names in published examples, and do not publish long transcripts in which the speakers could be identified.
- The invasion on privacy strongly depends on the context in which the recordings were made. For example, surreptitious recording of conversations in a public place, such as a bar, a supermarket, or department store, would be less invasive on the privacy of people than surreptitious recording of conversations in a bedroom. In public places speakers probably produce speech with the full awareness that some passerby may hear what is being said. As another example, surreptitious recording of telephone conversations will be more invasive on the privacy of people when they talk about confidential matters concerning various intimate details about their lives than when information about plane schedules is asked (as, for instance, in Wizard of Oz experiments). Also the age of the speakers who are being surreptitiously recorded may be important with respect to confidentiality; in general, the speech of (young) children will be regarded as less confidential than the speech of adults.
- And last but not least, it is strongly advised that people be informed afterwards that their speech was surreptitiously recorded and offer the opportunity to review and/or destroy all or part of the recordings. The argument for debriefing is perhaps somewhat less compelling when the researcher stays away from private conversations, but both from a legal and an ethical point of view debriefing would always be preferable.

A more elaborate discussion of legal and ethical aspects of surreptitious recording of speech can be found in Larmouth (1986) and Murray and Murray (1986).

5 SL corpus representation

5.1 Introduction

In this chapter the linguistic representation of spoken language corpora will be discussed. As stated in Chapter 3, one of the factors that determine whether a collection of speech is a speech corpus is the fact that the latter is augmented with linguistic annotation (i.e. a symbolic representation of the speech). Since it is impossible to examine the sampled speech data directly, it is only by means of the symbolic representation of the speech that one is able to navigate through the corpus. It is important to note that all types of representations of speech are the result of an analysis or classification of the speech. The representations are *not* the speech itself, but an abstraction from it. However, they are sometimes used as if they were the speech itself.

In most cases, the symbolic representation of the speech implies that a transcription of the speech is made. Transcriptions are used in many fields of linguistics, including phonetics, phonology, dialectology, sociolinguistics, psycholinguistics, second language teaching, and speech pathology. Transcriptions are also used in disciplines like psychology, anthropology, and sociology. The type of transcription very much depends on its purpose. In particular, this purpose determines the degree of detail that is required. For example, if a speech corpus has been designed to investigate the amount of time several speakers are speaking simultaneously in a dialogue, a very global transcription will be sufficient. If a corpus has been collected to establish differences in pronunciations of words, one needs to have a very precise segmental transcription.

Detailed phonemic or phonetic transcriptions of large scale spoken language corpora with many speakers and much (spontaneous) speech can never be achieved. This would be too time-consuming and expensive. Therefore most large speech corpora are provided with word for word transcriptions, i.e. word level orthographic representations of what has been said (e.g. the ATIS and Switchboard corpora). However, a medium sized corpus of read speech can be provided with a segmental transcription and even with labelling at the segmental level. Examples are the American English TIMIT corpus, which consists of 630 speakers each reading 10 sentences, and also the German PHONDAT corpora (1990 and 1992, both read speech) and German VERBMOBIL corpus (from 1993, spontaneous speech). An orthographic transcription (sometimes referred to as a *transliteration*) may be converted into a canonical phonemic transcription by means of a grapheme-phoneme converter or a pronunciation table.

It has been found that providing reliable phonetic transcriptions for large

corpora is hardly feasible (cf. Cucchiarini 1993). However, detailed transcriptions of a small number of specific phenomena (e.g. presence/absence of diphthongation, voiced/voiceless character of fricatives) can be made relatively fast and reliably if the occurrences of these phenomena can be retrieved quickly with the aid of annotation and direct access to files offered in a computerised speech corpus (cf. Van Hout 1989; Van Bezooijen and Van Hout 1985).

During the International Conference on Spoken Language Processing (IC-SLP) in Banff, Canada in 1992, a workshop was held on "Orthographic and Phonetic Transcription". The goal of the workshop was to agree on areas where community-wide conventions were needed, to identify and document current work, and to establish a means of future communication and continued cooperation.

In the remainder of this section some general remarks will be made about transcriptions of read speech versus transcriptions of spontaneous speech.

In addition, the levels and types of transcription will be introduced. In the next section (5.2), some background will be given on the task of segmenting and labelling speech. The following section (Section 5.3) will discuss the levels and types of representation in detail. For each level, reference will be made to existing corpora where possible, the symbols to be used will be presented, and recommendations will be given.

5.1.1 The transcription of read speech versus the transcription of spontaneous speech

The point of departure in the case of the transcription of read speech is the written text. This makes this type of transcription somewhat easier to perform than transcriptions of spontaneous speech, where an orthographic transcription must first be made. In the case of read speech, planning and word seeking processes are not involved. These processes of spontaneous speech production have a significant effect on the speech that is produced. It is well known that spontaneous speech is not fluent: speakers produce numerous filled pauses, mispronunciations, false starts, and repetitions. In addition, depending on the formality or informality of the setting, speakers will use colloquial speech and non-standard pronunciations. These properties of spontaneous speech make all types of transcriptions, global as well as detailed ones, more difficult to perform for spontaneous speech than for read speech. In the case of read speech, the use of written texts will ensure that there are fewer dysfluencies and a lower incidence of non-standard pronunciations.

Another important distinction between read and spontaneous speech in relation to transcriptions is that for read speech it is clear what an utterance is: the written sentence, usually starting with a capital letter and ending with a full stop. For spontaneous speech this is not necessarily the case.

Depending on the type of spontaneous speech involved, it is often necessary to define the criteria for delimiting utterances. For dialogues and other forms of conversation in which more than one speaker is involved, it is usual to define utterances more or less in terms of speaker turns (see the Guidelines issued by the Text Encoding Initiative (TEI) in Sperberg-McQueen and Burnard (1994) and Switchboard). For monologues, utterances can be defined as stretches of speech mostly preceded and followed by a pause and having a more or less consistent syntactic, semantic, pragmatic, and prosodic structure (see the criteria developed by the Network of European Reference Corpora (NERC) in French (1991, 1992) and the Dutch Speech Styles Corpus (Den Os 1994) (see also the results of the EAGLES Working Group on machine readable corpora).

5.1.1.1 Transcription of dialogues

When two (or more) persons are conversing together, interruptions frequently occur. (cf. Chapter 13). This is true for informal conversations between friends, for formal requests for information, for face-to-face situations, and for telephone conversations. These interruptions may be complete utterances, or they may be for instance affirmative "yes", or "mm". These interruptions in simultaneous speech must be annotated in the transcription. In the case of a dialogue between two persons, it is possible to give a clear indication of simultaneous speech. For example, Switchboard uses the "sharp" symbol ("#"), at either side of each of the simultaneous speek at the same time:

A: # Right, bye # B: # Bye bye #

In the case where more than two speakers are conversing, however, it is not possible to indicate the interruptions and simultaneous speech in a clear and simple way. For these cases a so-called "score notation" can be used. As for music score notation, the different speakers are given a separate track, one above the other. The tracks must be synchronised with one another. A computer program, known as "syncwriter", has been developed that handles this type of conversation, and runs on the Apple Macintosh.

It is also possible to collect dialogues that avoid simultaneous speech. In part of the VERBMOBIL corpus, the dialogue partners are recorded separately. The partners press a button when about to speak, which operates the recording procedure. The recordings are made in two rooms, separated by a glass screen so that the speakers can see each other. The speakers can hear each other by means of headphones. Clearly this situation is not as natural as the case where both speakers are permitted to speak at the same time.



"Bread and butter" (spoken in a Scottish accent) transcribed at four levels. Acoustic phonetic, Narrow phonetic levels: IPA symbols Broad phonetic, Citation phonemic levels: SAMPA symbols

Figure 5.1: Levels of annotation

5.1.2 Levels and types of transcription

All types of speech, whether read or spontaneous, whether monologue or dialogue, can be represented at different levels and in different ways. A distinction may be made between *transcriptions* and *representations that* can be derived from transcriptions. In making a transcription, it is necessary to monitor the actual speech. This can be done by the human ear or by means of a computer program. In the case of representations derived from transcriptions, it is not necessary to refer to the actual speech. These representations can be made at the phonemic level, and also at the morphological, syntactic, semantic, and pragmatic levels.

An orthographic transcription is often referred to as a "transcript" in courts of law (e.g. a transcript from a tape).

A distinction is made between the following types of annotation, which will be discussed in detail in the following sections. It is possible for all these levels of annotation to be time-aligned with the speech signal. However, time alignment is most likely to be used in the case of the prosodic, physical, acoustic-phonetic, narrow phonetic and (possibly) broad phonemic levels.

- 1. Recording script
- 2. Orthographic transcription
- 3. Morphological, syntactic, semantic, and pragmatic representations
- 4. Citation-phonemic representation
- 5. Broad phonetic transcription or phonotypic transcription
- 6. Narrow phonetic transcription
- 7. Acoustic-phonetic transcription
- 8. Physical transcription
- 9. Prosodic transcription

Figure 5.1 exemplifies the different levels, based on material in Barry and Fourcin (1992).

5.2 Concerning the segmentation and labelling of speech signals

Segmentation is the division of a speech file into non-overlapping sections corresponding to physical or linguistic units. Labelling is the assignment of physical or linguistic labels to these units. Both segmentation and labelling form a major part of current work in linguistic databases.

5.2.1 Background

The representation of a text or utterance as a string of symbols, without any reference to the acoustic form of the utterance, was the pattern followed by speech and text corpus work during the 1980s, such as the prosodicallytranscribed Spoken English Corpus (Knowles et al. 1995). These corpora did not link the symbolic representation with the physical acoustic waveform, and hence were not fully machine-readable. A recent project, MARSEC (Roach et al. 1993), has generated these links for the Spoken English Corpus such that it is now a segmented and labelled database.

The types of segments that may be delimited are of various kinds, depending on the purpose for which the database is collected. The German PHONDAT and VERBMOBIL-PHONDAT corpora use the CRIL (Computer Representation of Individual Languages) conventions, which propose three levels of representation: orthographic, phonetic and narrow phonetic.

A more detailed system of levels of labelling has been proposed by Barry and Fourcin (1992), which includes the above three levels. Each given speech corpus will choose one or more of these levels, which are described in detail in the following sections, and which grew out of the SAM project for the major European languages.

The format of label (transcription) files varies widely across research institutions. The WAVES format is becoming popular, and has the advantage of being human-readable. The recommendation is to use a label file format that can easily be converted to a WAVES label file, for the sake of portability across different systems.

5.2.2 A caveat

The treatment of speech as a sequence of segments to be delimited is to some extent a convenient fiction, made necessary by the requirements of speech technology. For example, it is notoriously difficult to define the boundaries between vowels and glides, or between a vowel and a following vowel. In addition, information about the place of articulation of a consonant is usually contained in its neighbouring vowels rather than the consonant itself. In the case of place assimilation, electropalatographic studies have shown that there is often a residual gesture towards the underlying segment (Nolan 1987). Hence one cannot describe the speech signal as a simple string of discrete phones in absolute terms.

Notwithstanding the above, Roach et al. (1990) argue that the attempt to segment speech is valid, as many segments (especially some consonants) have very clear acoustic boundaries. Where clear acoustic boundaries do not exist in the speech signal, selecting a fairly arbitrary point is better than doing no segmenting at all, from the viewpoint of speech technology research. Since segmented corpora may be useful for training HMM-based recognisers, problems of this kind could be cancelled out by including a great deal of data of the problematic kind, so as to avoid skewing the statistical models with only one view of the boundary location.

5.2.3 Use of segmented speech databases

A segmented and labelled speech database may be used for training the HMM models used in many recognisers, as well as for testing them. In addition, such a database provides the raw data needed for deriving rules for text-to-speech synthesis (rules of duration, intonation, formant frequencies, diphones, etc.).

A segmented and labelled speech database is also a primary resource in basic linguistic research, particularly in the case of little-researched languages. Such a database can yield fundamental information on the acoustic parameters of speech sounds of the language, as well as more detailed information on such things as patterns of duration variation according to linguistic context.

5.2.4 The CRIL conventions

The CRIL (Computer Representation of Individual Languages) conventions have been defined and proposed by a working group at the 1989 Kiel Convention of the International Phonetic Association. The conventions consist of two parts, as follows:
- 152 Part I: Spoken language system and corpus design
- 1. A mapping of IPA categories to standard IPA numbers and IPA symbol names. $^{\rm 1}$
- 2. Three distinct levels of representation for natural speech production.

The first component was introduced to enable broader use and dissemination of the descriptive IPA categories: i.e. the IPA symbols as well as the IPA diacritics needed for the narrow transcription of normal and defective utterances.

The second component of CRIL is devoted to a standardised representation of natural speech productions and introduces three systematically distinct levels for specifying what could be called the text of a spoken utterance. These levels are as follows:

1. Orthographic level

This level contains the orthographic representation of the spoken text.

2. Phonetic level

This level specifies the phonetic form of a given word in its full (unreduced) segmental form. This form would only appear when the word is spoken in isolation: i.e. it is the citation form.

3. Narrow Phonetic level

This level gives the narrow phonetic transcription of the words that were actually spoken. It is only on this level that phonetic categories can be directly related to the speech signal itself.

The CRIL conventions have been shown to work very well for the characterisation of speech data in the German PHONDAT and VERBMOBIL-PHONDAT corpora. However, the SAM conventions described below offer the opportunity to represent more detail.

5.2.5 Manual segmentation

Manual segmentation refers to the process whereby an expert transcriber segments and labels a speech file by hand, referring only to the spectrogram and/or waveform. There is no automatic assistance in segmenting. The manual method is believed to be more accurate. Also, the use of a human transcriber ensures that the segment boundaries and labels (at least at the narrow phonetic level) are perceptually valid. However, there is a need for explicit segmentation criteria to ensure both inter- and intra-transcriber consistency, together with (ideally) some form of checking procedure. Sets of guidelines for manual segmentation have been developed by various projects. One such is Hieronymus et al. (1990), which uses the four levels of underlying phonemic, broad phonetic, narrow phonetic and acoustic. It also retains the same base phonemic symbol even at the acoustic level, to facilitate the automatic determination of boundaries at the phonetic level once the boundaries at the acoustic level have been determined. Much speech data

¹ This item was not relevant for VERBMOBIL-PHONDAT, which made use of an extended SAMPA notation (7-bit ASCII).

(particularly in English) has been segmented and labelled entirely manually. This also applies to the spontaneous dialogue corpus in the VERBMOBIL project, part of which has been processed manually at IPDS Kiel (CD-ROMs 2,3; IPDS (1995, 1996)) on the basis of transcription conventions (modified SAMPA) laid down in Kohler et al. (1995).

One possible measure of accuracy for segmentation and labelling is consistency between transcribers. Barry and Fourcin (1992) quote Cosi and Omologo (1991) as saying that one should not expect more than 90% agreement between experts. Eisen (1993) investigates inter-transcriber consistency for the separate tasks of segmentation and labelling at three different levels of labelling, and concludes that consistency depends partly on the degree of abstraction of the labelling level, and partly on the type of sound involved. The best results in labelling were achieved at the broad phonetic level, for fricatives, nasals and laterals, which showed greater than 90% agreement across transcribers. The best results for segmentation were achieved at the acoustic-phonetic level, for the acoustic features "fricative", "voiced" and "vowel-like".

Recommendation 1

Any accuracy measure based on inter-transcriber consistency must control for the factors "level of transcription", "segment type", and "task type" (whether segmentation or labelling).

5.2.6 Automatic and semi-automatic segmentation

Automatic segmentation refers to the process whereby segment boundaries are assigned automatically by a program. This will probably be an HMMbased speech recogniser that has been given the correct symbol string as input. The output boundaries may not be entirely accurate, especially if the training data was sparse. Semi-automatic segmentation refers to the process whereby this automatic segmentation is followed by manual checking and editing of the segment boundaries.

This form of segmenting is motivated by the need to segment very large databases for the purpose of training ever more comprehensive recognisers. Manual segmentation is extremely costly in time and effort, and automatic segmentation, if sufficiently accurate, could provide a short cut. However, it is still necessary for the researcher to derive the correct symbol string to input to the autosegmenter. This may be derived automatically from an orthographic transcription, in which case it will not always correspond to the particular utterance unless manually checked and edited. The amount of inaccuracy that is acceptable will depend on the uses to which the database is to be put, and its overall size.

There will always be a need to verify the accuracy of an autosegmented database, and the obvious accuracy measure is the consistency between

manual and automatic segmentation over a given subset of the database. Schmidt and Watson (1991) carried out this evaluation over nearly 6000 phoneme-sized segments, and found that the discrepancy between manual and automatic boundaries varied across segment types. The absolute mean discrepancy was greatest for diphthongs (5.4 ms) and least for nasals (0.37 ms). For 50% of all segmentations, the discrepancy was less than 12 ms, while for 95% it was less than 40 ms. This falls within the range of just-noticeable differences in duration for sounds of the durational order of speech sounds (Lehiste 1970, p. 13) and so one could conclude that the discrepancies are not perceptually relevant. This means that automatic segmentation for the given data, using the given autosegmenter, was probably sufficiently accurate.

5.2.7 Segmentation and labelling in the VERBMOBIL project

In the German VERBMOBIL project, segmentation and labelling of recorded speech data is a fundamental part of the research. The following procedures are adopted (Kohler et al. 1995). The phonemic labels are based on the SAMPA symbols for German, augmented by extra labels for phonetic segments such as plosive release, aspiration after a plosive, creak, and nasalisation of a vowel as a reflex of a deleted nasal. Hence the labelling is partially carried out at the narrow phonetic level, being basically phonemic. During the labelling process, a label will be aligned with the start of the portion of speech that is considered to represent its chief acoustic correlates. Labels are discrete and non-overlapping, except in the following cases:

- 1. Labels for creak and nasalisation are always superimposed on other labels, which they modify.
- 2. A special label (-MA) is used to indicate that the phonetic correlates of one or more deleted segments are present in the surrounding material. For example, where an unstressed rounded vowel has been deleted, labialisation may still be present in a neighbouring consonant, and will be marked in this way.

Inter-labeller consistency is maintained in three ways, as follows:

- 1. The inventory of possible labels is restricted mainly to the list of German phonemes. This restriction minimises the possibility of error.
- 2. The labeller works from a citation-phonemic form of the utterance that has been previously prepared. This eliminates gross errors.
- 3. There are restrictions on the types of modification allowed. The labeller is permitted to mark the following: deletion (where the initially-provided label is marked with a following hyphen, to indicate deletion); insertion (where the new label is prefixed with a hyphen); and substitution (where the new label is inserted after the one initially provided, separated by a hyphen).

The checking of segmented and labelled speech files is carried out partly

by a program developed at IPDS Kiel that detects invalid sequences of symbols, and partly by experienced labellers checking the work of less experienced transcribers (Kohler et al. 1995). All segmenting and labelling is carried out manually. The initial citation-phonemic transcription is generated at IPDS Kiel with the help of the grapheme-to-phoneme converter of the Rulsys/Infovox TTS system for German (Kohler et al. 1995), subsequently checked manually for mistakes. A system of prosodic labelling has also been developed: PROLAB (Kohler et al. 1995).

5.3 The levels and types of symbolic representations of speech

5.3.1 Recording script

In the case of speech read from a script, the simplest and quickest level of annotation is the orthographic form of the words in the recording script, as these are readily available. However, most researchers will need more detail than this, and so other levels must also be considered.

5.3.2 Orthographic transcription

By orthographic transcription is meant the use of the standard spelling conventions of the language. Orthographic transcriptions are used in large scale speech corpora and in corpora used for research in which details about the pronunciation of words are not important. In the first case, a detailed transcription of at least part of the corpus is perhaps desirable (especially if such a corpus is used for training of speech recognition systems), but because of the huge amount of work this cannot always be done. In the latter case, precise transcriptions are simply not necessary.

The use of normal spelling necessarily implies a compromise between the sounds heard and what is written down. Particularly in the case of spontaneous speech, there may be a significant discrepancy between what is heard and the symbolic representation used to encode it (see Appendix M for the guidelines for orthographic transcription used in the SPEECHDAT corpora).

5.3.2.1 Reduced word forms

Because of the discrepancy between what is heard and what is written, many developers of spontaneous speech corpora have decided to provide an indication of reduced word forms, using the reduced forms given in the standard dictionary for the language. However, in the interests of consistency, developers are sometimes forced to use forms not present in the dictionary. For example, in German, a preposition is often contracted together with a following article, forming one word and thus reducing the number of syllables (e.g. "zu der" is pronounced and written as "zur"). These forms occur in the Duden dictionary (a standard German dictionary). However, in the VERBMOBIL corpus, the developer is also permitted to write "fürn"

for "für den", although this form does not exist in the dictionary. Similarly, in the Dutch Speech Styles corpus it was decided to indicate reduced word forms. Criteria for indicating reduced forms in an orthographic transcription may be a) frequency of occurrence of these forms and b) reduction in the number of syllables. The reduced word forms used should be listed in the accompanying documentation.

5.3.2.2 Dialect forms

Even in speech corpora covering the standard variety of a given language, speakers may have their own idiolect or may use words that have a dialect basis. These words have to be marked in the transcription. The developers of the VERBMOBIL corpus chose an orthographic means of indicating dialect words which are not in the Duden dictionary. It is possible to give information about the meaning behind these words, as in the following, explaining the dialect form of the greeting "good morning" spoken in the North of Germany: "moin, moin <; norddeutsche Grußformel>"

5.3.2.3 Numbers

In orthographic transcriptions, numbers are usually spelled out in full rather than being written in digit form. In some cases, the decision is made to deviate from the standard spelling in order to avoid excessively long words. For example, in the VERBMOBIL corpus the numbers 13 to 99 as well as the hundreds from 1 to 19 (the years) are written as a single word in accordance with German orthographic conventions. However, all other numbers are written separately, and thus do not conform to the normal rules. Examples follow:

1993:	neunzehnhundert dreiundneunzig					
3049614:	drei Millionen neunundvierzig tausend					
	sechshundert vierzehn					
349:	dreihundert neunundvierzig					

5.3.2.4 Abbreviations and spelled words

In orthographic transcriptions, the full form of an abbreviation is usually written. Hence "e.g." is written as "for example", and German "usw." is written as "und so weiter". Abbreviations which are pronounced as words in their own right are spelled as words (e.g. Benelux, OPEC, NATO). In spoken language corpora such as POLYPHONE, speakers were asked to

read out (among other things) spelled words. Words can be spelled out in different ways, as follows. Firstly, the names of the letters can be pronounced (e.g. A, B, C), but one can also use words beginning with the letter concerned, like Alpha, Bravo, Charlie (as in the radio alphabets used by the military etc.). Spellings must be indicated in orthographic transcriptions, including the case when only part of the word is spelled out, as in (German) "USA-trip", "Vitamin-C". In the VERBMOBIL corpus, spellings are indicated by capitals preceded by \$: \$U-\$S-\$A-trip, Vitamin-\$C.

5.3.2.5 Interjections

Interjections such as "ah", "oh", "mm", or the French "hein" must be shown according to the standard spelling of these forms in the given language. If there is no standard spelling for a certain interjection, it is necessary to decide on a spelling, and ensure it is included in the documentation associated with the corpus.

5.3.2.6 Orthographic transcription of read speech

When orthographic transcriptions are used for corpora containing read speech, the original written text may function as the default transcription. The transcription indicates how well the written text was read by the speaker. For single words or short sentences, speakers will make relatively few mistakes, as has been found by the Dutch POLYPHONE Corpus and the GRONINGEN Corpus. However, in the case of read texts (even short texts), it appears that speakers often make mistakes. These mistakes are mostly related to deletions of words, false starts, or hesitations. Speakers may also add words not present in the text, or they may use a different word order. Furthermore, speakers may mispronounce words. For example, they may add, omit, or scramble syllables.

Depending on the intended application of a corpus containing read as well as spontaneous speech, such dysfluencies must be indicated in the orthographic transcription. If the corpus is to be used for initialising speech recognition systems, every sound must be annotated, including hesitations, filled pauses etc. Research on reading errors will also require annotation of such dysfluencies. On the other hand, if a corpus is to be used to determine the type of syntactic structures typically used in a certain dialogue system, then it is not necessary to indicate all events occurring in the signal. For an overview of annotations which can be used in an orthographic transcription and subsequent phonetic labelling see Kohler et al. (1995). These annotations concern verbal sounds made by the speaker (e.g. hesitations, coughing, laughing), as well as background noises (slamming of doors, ringing of telephones).

5.3.2.7 Orthographic transcription as the first of many levels

Some projects may wish to produce the orthographic transcription as the first of several linguistic levels of annotation. In this case, the next level will be the citation-phonemic form of the speech. If an on-line

pronouncing dictionary is available, the citation-phonemic form may be derived automatically from the orthography, thus saving time.

Recommendation 2

When transcribing a corpus orthographically, it is advisable to generate a list of all unique word forms found in the transcription. This list will then form the input to a grapheme-to-phoneme conversion module (which may involve accessing a phonemic dictionary and/or running letter-to-sound rules). The output of this module will be a table with the citation-phonemic forms (canonical forms) of the speech, which can form a basis for later adaptation to various accents of the same language. This procedure is followed, for example, in the SPEECHDAT corpora.

5.3.3 Morphological, syntactic, semantic and pragmatic representations

The database may also be labelled or "tagged" at linguistic levels above the phoneme, such as the morphological, syntactic, semantic and pragmatic levels. This is of particular interest to those wishing to study, for example, the relationship between prosody and syntax. On the syntactic level, the researcher will label such items as clause and phrase boundaries, compound word internal boundaries, and possibly affix boundaries also. These units may be derived from the orthographic transcription, and so there is no need to examine the speech waveform for this level of representation. Automatic tagging software developped for large written text corpora may be used here, although adaptations are necessary for spontaneous speech.

5.3.4 Citation-phonemic representation

The citation-phonemic level (also referred to as the "phonemic" level by Barry and Fourcin (1992) may contain the output phoneme string derived from the orthographic form (by lexical access, by letter-to-sound rules, or both). There are various possibilities for representing the phoneme symbols. Some platforms have the facility to display the full range of IPA symbols, such as the symbols used by the IAT_FX font wsuipal1 (see Table 5.1).

However, many researchers will need to use some other means, such as an alphabetic or numeric representation of IPA symbols. The numeric equivalents of all IPA symbols are displayed in Esling and Gaylord (1993). An alphabetic equivalent system is used on the newsgroup sci.lang.

If the requirement is narrowed to symbols only for the main European languages, then the SAMPA system (see Appendix B) will be sufficient. This system has the advantage that, for any given language, only one grapheme is used per phoneme, with no spaces in between. Other systems that have been proposed (principally for English) allow a string of two or more graphemes to represent a phoneme, but a space between each phoneme representation is then necessary. In the case of English, there are still more alphabetic systems that have been used in the past, such as (for American English) ARPABET and KLATTBET, and (for British English RP) Edinburgh's Machine-Readable Phonetic Alphabet, all of which use short grapheme strings separated by a blank space. However, a language-specific set of alphabetic phoneme symbols has not yet been devised for all possible languages.

RECOMMENDATION 3

If the corpus is confined to one language, and if the labelling is to be alphabetic rather than true IPA symbols, then it is advisable to use a languagespecific set of characters. This avoids the notational complexity necessary when all symbols must be kept distinct across all languages, as is needed in the study of general phonetics.

5.3.5 Broad phonetic or phonotypic transcription

The citation-phonemic representation will not show such phenomena of running speech as place assimilation, consonant deletion, French liaison, RP linking-r and vowel reduction, etc. Hence there is a need for this more detailed level, which may (at least initially) be generated entirely by phonological rules from the citation-phonemic representation. It uses only symbols that have the status of phonemes, marking the output of connected speech processes that either insert or delete phonemes, or transform one phoneme into another.

This is the level referred to as "phonotypic" for French labelling by Autesserre et al. (1989). It may be derived initially from the citationphonemic labelling by phonological rules which, for example, reduce or delete unstressed vowels in English, or which delete some cases of word-final schwa in French. At this stage, it does not require reference to the sounds actually produced by the speaker (though this may come later, in manual post-editing). The "broad phonetic" level of Barry and Fourcin (1992) is intended as a phonemic-level representation of the speaker's tokens, i.e. the transcriber makes reference to the speech signal. The "phonotypic" level of Autesserre et al. (1989) is intended to be derived purely by rule. Both approaches are possible: however, there are certain advantages in deriving this level of representation purely by rule, as this is relatively quick and easy. This level uses a limited inventory of symbols (and so it is still a practical proposition for very large databases), while also offering more phonetic detail than the citation forms. So there is a balance between accuracy of representation and ease of derivation of the representation. However there will still be discrepancies between completely rule-derived labelling of an automatically segmented corpus) and the utterance, and hence some loss of accuracy; this may be offset by a large volume of data.

5.3.6 Narrow phonetic transcription

The narrow phonetic level of representation is the first one where the transcriber cannot avoid listening to the recording and/or inspecting the waveform and spectrogram. This is because it attempts to represent what the speaker actually said at the time of recording. This consideration immediately increases the time and effort necessary, as every part of the speech must be inspected manually.

The inventory of symbols is increased to include sounds that do not have phonemic status in the language (such as a glottal stop for English, or a symbol for aspiration after a voiceless plosive). It is at this level that different allophones may be represented, as well as devoicing or voicing, and secondary articulations such as nasalisation or labialisation. One segment at this level (e.g. a voiceless plosive) may correspond to more than one segment on the acoustic-phonetic level (e.g. the closure phase and burst of a plosive). There are potential problems in determining the boundaries of segments at this level, especially in the case of "desequencing" or transposition of two sounds (Barry and Fourcin, op. cit.). However, this level of representation will be far more accurate as a record of what was said. This implies that not so much data will be needed, provided it is accurately transcribed at this level.

Recommendation 4

It is better not to embark without good reason on this level of representation, which requires the researcher to inspect the speech itself, as this greatly increases the resources needed (in terms of time and effort). If the broad phonetic (i.e. phonotypic) level is considered sufficient, then labelling at the narrow phonetic level should not be undertaken.

5.3.7 Acoustic-phonetic transcription

This level of representation distinguishes every portion of speech that is recognisably a separate segment of the acoustic waveform or spectrogram. Hence stop closures, release bursts, and aspiration phases will all be segmented separately, as will glottal onset of a vowel, or separate voiced and voiceless portions of a fricative, nasal or stop. The labelling will be done in terms of these well-established articulatory categories, which can be easily related to the labels at higher linguistic levels. The boundaries of some segment types (glides and vowels) will to some extent be arbitrary, and a criterion for homogeneity such as "sameness of change" will have to be employed (e.g. in the falling second formant for a palatal glide).

5.3.8 Physical transcription

The most detailed level of representation is the physical level. This does not need to relate only to an acoustic record, but could have separate tiers related to different types of input (e.g. nasal transmission detectors, palatography). However, acoustic parameters are likely to be the most frequent type of representation, and different types may be needed for particular application areas (e.g. filter bank output energies, formant frequencies, LPC and cepstral coefficients, fundamental frequency or electroglottograph output waveform). The physical events may be overlapping or discrete in time, with each parameter allotted a separate annotation row (e.g. nasal resonance, periodicity, high-frequency noise). This level of labelling has not been generally used in speech technology research to date, but it has the potential to serve as a resource for developing speech synthesisers with greater "naturalness", and speech recognisers which include more speech knowledge in their algorithms.

5.3.9 Prosodic transcription

It is possible to mark some prosodic information even in orthographic transcriptions, such as lengthening of sounds, pauses in words and utterances, emphatic stress, and intonational boundaries. Examples are the ATIS and Switchboard corpora, and the Dutch Speech Styles Corpus. If more detail than this is required, however, it is necessary to undertake a full prosodic transcription. For an overview of some existing prosodic transcription approaches, see Llisterri (1994). The following will describe prosodic transcription in general terms.

5.3.9.1 Types of approach to prosodic labelling

The above discussion has been in terms of segmental labelling only. It is also possible to annotate a speech database at the prosodic (suprasegmental) level. This is less straightforward than segmental annotation, as there are far fewer clear acoustic cues to prosodic phenomena. The F0 curve will be the relevant acoustic display, possibly augmented by the intensity curve. The waveform is a useful guide to the current location in the speech and is usually displayed together with the F0 curve (as in the WAVES labelling software).

The units segmented will depend on the particular theoretical bias underlying the given research programme. A basic distinction may be drawn between a prosodic labelling system that annotates the boundaries of units (analogous to the method used in segmental annotation) and a system that annotates the occurrence of isolated prosodic events, such as F0 peaks.

The first type of method may possibly use the intonational categories proposed by Nespor and Vogel (1986), such as intonational phrase, phonological

phrase, phonological word, foot, and syllable. Alternatively, it could mark the more traditional units of "minor tone-unit" and "major tone-unit", as in the MARSEC database (Roach et al. 1993). Or again, it could annotate the perceptual phonetic categories used in the "Dutch school" of intonation studies, such as rises and falls that are early, late or very late in their timing, fast or slow in their rate of change, and full or half sized ('t Hart et al. 1990). This type of annotation could be used in conjunction with annotation at the morphosyntactic level to yield information about the relationship between the syntactic and prosodic levels in terms of duration, pauses, etc.

The second type of method, though it may refer to the units mentioned above in its underlying theory, does not in fact annotate them but rather marks the occurrence of high and low tones of various kinds. The recently formulated ToBI transcription system (Silverman et al. 1992) is the most well-known system of this kind for English, where the prosodic units are annotated at the "break index" level rather than the "tone" level. (For an account of prosodic labelling for German see Reyelt et al. (1996)). Other systems, such as SAMPROSA (see Appendix B) have also been proposed.

5.3.9.2 Examples of the two types of approach

Prosodic annotation has only recently come into favour in the field of speech and language technology research. Now that a basic level of competence has been achieved as regards the synthesis and recognition of speech segments, researchers have come to realise that much more work is required on the prosodic aspect of speech technology. This is the motivation for the growth in popularity of speech database research, and for the formulation of the ToBI prosodic transcription system. In order for the prosodic transcriptions of various different speech databases to be comparable, and in order to make the best use of existing resources, the originators of ToBI (Silverman et al., op. cit.) proposed a simple system that would be easy to learn and that would lead to good inter-transcriber consistency. To date it has largely been used for English, especially American English, but at least in principle it could be extended to other languages as well. The system has certain severe limitations (e.g. it has no way of representing pitch range) but its minimalist formulation was dictated by the need for learnability and consistency in use. The "British school" type of system used in the MARSEC database of British English (Roach et al. 1993) contains more phonetic detail but may require more effort in teaching to novice transcribers. The "IPO" classification of F0 patterns ('t Hart et al. 1990) has not yet been used systematically in the annotation of large-scale publicly-available speech corpora, but has been used successfully in the development of speech synthesisers.

Prosodic transcription also has obvious uses in basic linguistic research, especially since research into the suprasegmental aspects of language is not nearly as advanced as research into the segmental aspects. As indicated

above, a database annotated at the prosodic and morphosyntactic levels can provide information on the relationship between them with respect to duration and pauses. If the segmental level is also annotated, then many possibilities open up for the study of segmental duration in prosodic contexts. This is especially true in the case of languages other than English, where these aspects have received comparatively little attention to date. The concept of levels of prosodic labelling applies differently to the two different approaches to prosodic labelling outlined above. In the first case, the obvious categories would be those proposed by Nespor and Vogel (op. cit.), comprising levels of non-overlapping units each of which corresponds to one or more units on the level immediately below (e.g. phonological phrase, foot, syllable). In the second case, the separate levels have no such intrinsic relationship to one another, but merely deal with different types of phenomena. For example, in the ToBI system, there are separate levels for tones and inter-word "break indices". The ToBI system can be described briefly in terms of its separate levels, and is described below. The MARSEC system will be outlined after that. The "Dutch school" system of IPO will not be described in much detail, as it has not yet been used for annotation of publicly-available speech corpora: however, extensive references are available in 't Hart et al. (op. cit.).

5.3.9.3 The ToBI labelling system

A recent experiment (Pitrelli et al. 1994) used several prosodic transcribers working independently on the same speech data, comprising both read and spontaneous American English speech. The ToBI system was used, and a high level of consistency across transcribers was found, even given the fact that transcribers included both experts and newly-trained users of the system. This suggests that the system achieves its object of being easy to learn and to apply consistently, at least in the case of American English.

The "orthographic" level of the ToBI system contains the orthographic words of the utterance (sometimes only partial words in the case of spontaneous speech). It is also possible to represent filled pauses (e.g. "um", "er") at this level.

The "miscellaneous" level may be used to mark the duration of such phenomena as silence, audible breaths, laughter and dysfluencies. There is no exhaustive list of categories for this level, and different transcription projects may make their own decisions as to what to annotate.

The "break index" level is used to mark break indices, which are numbers representing the strength of the boundary between two orthographic words. The number 0 represents no boundary (with phonetic evidence of cliticisation, e.g. resyllabification of a consonant), and 4 represents a full intonation phrase boundary (usually "end of sentence" in read speech), defined by the occurrence of a final boundary tone after the last phrase tone. The number

3 represents an intermediate phrase boundary, defined by the occurrence of a phrase tone after the last pitch accent, while the number 1 represents most phrase-medial word boundaries. The number 2 represents either a strong disjuncture with pause but no tonal discontinuity, or a disjuncture that is weaker than expected at a tonally-signalled full intonation or intermediate phrase boundary.

The "tone" level is used to mark the occurrence of phonological tones at appropriate points in the F0 contour. The basic tones are "L" or "H" (for "low" and "high"), but these may function as pitch accents, phrase accents or boundary tones, depending on their location in the prosodic unit. In the case of pitch accents (which occur on accented syllables), there may be one or two tones, and the H tone may or may not be "downstepped".

Information about the ToBI system and guidelines for transcribing are available on the Internet.

5.3.9.4 The MARSEC labelling system

The MARSEC project (Roach et al. 1993) is based on the Spoken English Corpus (Knowles et al. 1995), a corpus of British English that at the time was not time-aligned. The MARSEC project time-aligns the prosodic annotations, the orthographic words, the grammatical tag of each word, and individual segments. The type of prosodic annotation used is the "tonetic stress mark" type of system. Several types of accent are recognised: low fall, high fall, low rise, high rise, low fall-rise, high fall-rise, low rise-fall, high rise-fall, low level, and high level. These may occur either on nuclear or on non-nuclear accented syllables. In addition, there is a distinction between major and minor tone-unit boundaries, and there is provision for "markedly higher" or "markedly lower" in perceived pitch. The tonetic stress mark type of system has been used for many years, and has been applied to many languages apart from English (the same is not true of the ToBI system). However, no extensive attempts have yet been made to apply it in the field of speech technology.

The Spoken English Corpus comprises over fifty thousand words of broadcast British English in various styles, mostly monologues. Two transcribers prosodically annotated it in an auditory fashion, with no access to the F0 curve. They each transcribed half the corpus, but each also independently transcribed certain passages known as "overlap" passages, the purpose of which was to check on inter-transcriber consistency. Analysis of the overlap passages reveals that the consistency is fairly good, certainly in the case of major aspects such as location of accents and direction of pitch movement (Knowles and Alderson 1995). This result is especially encouraging in view of the fact that the transcription system used contains far more phonetic detail than does the ToBI system.

5.3.9.5 The IPO approach

The phonetically-based analysis of intonation used at IPO ('t Hart et al. 1990) has the advantage of having proved its usefulness for more than one language, and of having been successfully applied in the field of speech synthesis (neither of these considerations apply to the ToBI system). The analysis proceeds by modelling F0 curves in terms of straight lines that have been experimentally proved to be perceptually indistinguishable from the original ("close-copy stylisations"). This type of representation is then further simplified into "standardised stylisations" in terms of a small set of available contours for a given language. This type of representation has been experimentally proved to be distinguishable from the original on close listening, but yet not functionally any different from the original (i.e. the standardised stylisation is linguistically equivalent).

In the case of Dutch, there are ten basic pitch movements (the model has also been applied to British English, German and Russian). These are the five falls and five rises, varying along the parameters of syllable position, rate of pitch change, and size of pitch excursion. These ten pitch movements are grouped into "pitch configurations" (of one or two pitch movements each). The pitch configurations are classified into prefixes, roots and suffixes. Sequences of pitch configurations are grouped into valid "pitch contours", which in turn are grouped into melodic families or "intonation patterns" (of which there are six in Dutch). These groupings are experimentally verified by listeners. The units of this analysis, at all levels, are based on speech corpora of spontaneous and semi-spontaneous utterances in Dutch. In contrast to the ToBI and MARSEC systems, comparatively little effort has been put into checking inter-transcriber consistency, possibly because the detection and labelling of this kind of phonetic unit is less problematic.

5.3.9.6 Prosodic labelling in the VERBMOBIL project

In the VERBMOBIL project, a large database of German spontaneous speech is being recorded at Munich, Bonn, Kiel and Karlsruhe. It covers a variety of different German speaking styles. Part of these data are being prosodically labelled at IPDS Kiel according to the VERBMOBIL prosodic conventions PROLAB (Kohler et al. 1995). Another section of the corpus is being processed at Braunschweig University according to an adapted ToBI system, along the following guidelines:

- 1. The data should be usable by several project partners for a variety of purposes. Therefore the inventory of labels was designed at several workshops held with potential users of the data.
- 2. It is necessary to use transcribers with little experience in prosodic labelling, in order to transcribe the amount of data needed for the training of statistical speech recognisers.

The tasks include not only the labelling of speech data, but also the development of a workstation for prosodic labelling, and methods and tools for increasing labelling speed and consistency, as follows:

- Segmentation of word boundaries is assisted by an automatic presegmentation algorithm using a SCHMM forced recogniser.
- Potential phrase boundaries are predicted by a statistical linguistic analysis.
- Resynthesis of the original speech signal is carried out with variation of F_0 according to the pitch accents labelled. This is done for the sake of verifying the transcription).
- Special-purpose software ("fish") is used for visualisation and labelling. The software uses Tcl/Tk, is public domain and runs under most versions of UNIX.
- Data exchange is carried out using the SAM data format.
- Evaluation of the labelling is carried out, checking for consistency and the influence of tools)

The label inventory splits into three tiers, as follows:

- 1. Functional tier: main accent, secondary accent, emphasised/contrastive accent, sentence modality.
- 2. Break index tier: (full) intonation phrase boundary, minor boundary, irregular boundary.
- 3. Tone tier: pitch accents and boundary tones.

The functional tier provides information about prosodic function like focus and modality.

- ? *Question mark*: Several question types are labelled. The recognition of questions and non-questions is one of the disambiguation tasks of prosody in VERBMOBIL.
- PA *Main accent*: In each intonational phrase the *most prominent word* is labelled. There can be more than one PA per intonational phrase, however.
- NA Secondary accent: All other accents are secondary accents.
- EK *Emphasised or contrastive accent*: This is an emphasised (i.e. extraordinarily strong) accent.

This functional tier seems to be unique among labelling systems. The reasons for the introduction of an explicit functional tier are as follows:

- 1. For a lot of users, this tier is the most important one besides the break index tier. The linguistic levels in speech recognition tend to use this functional part directly, while other systems merely need such information as unaccented/accented/strong accent.
- 2. The functional tier can also be found implicitly in other labelling systems. However, in the case of spontaneous speech, it appears to be more appropriate to mark the prominence relations explicitly, rather than assuming them from accent and boundary positions.

- 3. Focus information is very interesting in spontaneous speech. Although phrase accent and focus are not identical, the phrase accent may give clues about the location of the focus (see the similarities between the Stuttgart tones and the Braunschweig PA).
- 4. Previous experiments have shown that even inexperienced transcribers can mark phrase accents with a high degree of inter-transcriber consistency.

The break index tier marks different types of word boundary, as follows:

- B1 Normal word boundary.
- B2 Minor (intermediate) phrase boundary. This is a weak intonational marking, comparable to the "trail" tones used at the University of Stuttgart, and possibly to a "3" boundary in the ToBI system.
- B3 Full intonational boundary. This is a strong intonational event either with or without lengthening or change in speech tempo.
- B9 Irregular boundary. This boundary marks dysfluencies at hesitations, repairs etc.

The tone tier uses a ToBI-like inventory consisting of H and L tones (see also Reyelt et al. 1996). The pitch accents and boundary tones are intended as a phonologically distinctive minimal system, together with additional distinctions which proved to be necessary for labelling spontaneous speech. The accents are as follows:

H* This is the "normal" peak accent. It occurs as a rising accent in prenuclear position, and a falling accent in nuclear position before a low boundary tone.

The auditory impression within the accented syllable is "high".

 $L+H^*$ This is a medium (or rather a "raised") peak. Beginning with a low tone, the F_0 rises to a high peak within the syllable. This accent type is often combined with emphasis.

The auditory impression within the accented syllable is "high".

- L^*+H This is a "delayed" peak, with a H^* accent that reaches high F_0 in the syllable after the accented one. The German term for it is "Schleifton". The auditory impression within the accented syllable is "rising, between low and high".
- $\rm L^*~$ This is a "trough accent". It can take a rising form when followed by a H-H% boundary.

The auditory impression within the accented syllable is "low".

- !H*, L+!H*, L*+!H These are downstepped accents.
- $\rm H+!H^*~$ This is an "early peak". It comprises a fall before the accented syllable, and is often followed by a low boundary.

The auditory impression within the accented syllable is "low".

Each intonational phrase boundary is marked by two tones, a phrase tone and a boundary tone. These are both labelled, even if there is no clear bitonal pitch movement (and especially at low boundaries).

- 168 Part I: Spoken language system and corpus design
- L-L% This is a "terminal fall", where the F_0 reaches the bottom of the speaker's pitch range.
- H-H% This is a "question/continuation rise", where the F_0 approaches the top of the speaker's pitch range. There is a very steep rise after a low nuclear accent, or a high plateau after a H* accent where the high tones cause F_0 to remain high.
- L-H% This is a low phrase tone with a pitch rise to the mid or high level. This rise must be distinguished from a general slight final rise that is idiomatic for some speakers, and is discarded.
- $\rm H\text{-}L\%$ This is a "continuation fall". In some cases it involves a fall from high to mid pitch, and in others a more level boundary at mid pitch.

5.3.9.7 Provisional recommendations for prosodic transcription

It is reasonable to assume nowadays that a prosodic transcriber will have access to at least the waveform and the F0 curve for the speech to be transcribed. In that case, the recommendation is to use either the ToBI or the IPO system (and the MARSEC system if a purely auditory transcription is being carried out). If the language to be transcribed is not English, and especially if the projected application of the prosodic transcription is in the field of speech technology, then it is probably best to use the IPO system if possible (i.e. if the basic "grammar" of contours has already been researched for that language). However, these can only be provisional recommendations, as little work has been carried out on prosodic labelling in comparison with the great effort that has been expended on segmental labelling. In this situation, it may be that a different system entirely will prove more appropriate to the given language, and it is not possible to make absolute recommendations.

5.3.10 Non-linguistic and other phenomena

Another, additional level of representation may be used for non-linguisitc phenomena which occur when people are speaking. This includes speaker noises such as coughing, laughter, and lip smacking, as well as extraneous noises such as the barking of dogs and the slamming of doors. In addition, this level can also be used to label information such as dysfluencies and filled pauses. The type of representation used for such annotations will depend on the purpose of the database. An annotation system such as that proposed by the Text Encoding Initiative is very elaborate and makes heavy demands on a transcriber, but also makes it possible to derive all relevant information from a transcription. While the TEI system makes use of SGML, which guarantees that existing software can be used, there is a large initial learning curve for the transcriber, which multiplies the possibility of human error in the transcription. Other annotation systems (such as those used in ATIS and Switchboard) are less elaborate, but also easier for transcribers to learn. The conventions used in ATIS, Switchboard, POLYPHONE and the GRONINGEN corpus consist of different types of brackets with possible additional glosses. Retrieval software referring to these particular annotations must be designed in a more or less ad hoc way, which is less convenient than the TEI system. However, it is possible to provide standard UNIX scripts for a speech corpus. It is important to find the correct balance between the sophistication of the annotation system and the practicality of the system from the transcriber's point of view.

The types of phenomena which could conceivably be annotated on this level of representation are listed below.

1. Omissions in read text

Words from the recording script which were omitted by the speaker may be indicated. In spontaneous speech, it is very difficult to know whether a speaker has omitted words which he actually intended to say, and so omission is only relevant in the case of read speech.

2. VERBAL DELETIONS OR CORRECTIONS, IMPLICIT OR EXPLICIT

Words that are verbally deleted by the speaker may be indicated. Verbal deletions are words that are actually uttered, then (according to the transcriber) superseded by subsequent speech. This can be done explicitly, as in *Can you give me some information about the price*, *I mean*, the *place where I can find* ... Alternatively, it can be done implicitly, as in *Can you give me some information about the price*, *place where I can find* Verbal deletions or self-repairs may be indicated in read as well as spontaneous speech.

3. Word fragments

Word fragments comprise one or more sounds belonging to one word. For example, in ATIS word fragments are indicated by a hyphen, as in *Please show fli-flights from Dallas*.

4. UNINTELLIGIBLE WORDS

Sometimes only part of a word is unintelligible, in which case only the intelligible part is transcribed orthographically. If a word is completely unintelligible, that fact will be annotated on this level. For example, by putting "[unintelligible]" in the text (ATIS), or by putting two stars "**" as in SPEECHDAT corpora.

5. Hesitations and filled pauses

Filled pauses (such as uh and mm) may be indicated. Some annotation conventions (e.g. POLYPHONE and Switchboard) annotate only one or two types of filled pause (uh and mm, or only uh). Other systems (e.g. ATIS and Speech Styles) annotate more than two types (e.g. uh, mm, um, er, ah). The types of filled pause vary across languages (for example, the British English er is not used in Dutch). The recommendation is to use at least two types: one vowel-like type uh, and one nasal type mm.

6. NON-SPEECH ACOUSTIC EVENTS These can be made either by the speaker or by outside sources. The first

category includes lip smacks, grunts, laughter, heavy breathing and coughing. The second category includes the noise of doors slamming, phones ringing, dogs barking, and all kinds of noises from other speakers. The Switchboard corpus uses a very extensive list of non-speech acoustic events, ranging from bird squawk to groaning and yawning. The recommendation is that these events are annotated at the correct location in the utterance, by first transcribing the words and then indicating which words are simultaneous with the acoustic events.

7. Simultaneous speech

For dialogues and interviews, words spoken simultaneously by two or more speakers may be indicated.

8. Speaking turns

Discourse analysis makes use of indications of different speaking turns and initiatives. While these are not generally used in speech technology, it would always be possible to transcribe them.

5.4 List of recommendations

- 1. For the transcription of dialogues between more than two speakers use a "music score notation".
- 2. For orthographic transcriptions, use the standard spelling as much as possible.
- 3. Indicate reduced word forms in orthographic transcriptions a) if these forms occur frequently and b) if they involve syllable deletion.
- 4. Use at least two types of "filler" syllable: one vowel-like type uh , and one nasal type mm .
- 5. Non-speech acoustic events should be annotated at the correct location in the utterance, by first transcribing the words and then indicating which words are simultaneous with the acoustic events.
- 6. When orthographic transcription is used in a corpus, it is recommended that a list of unique words and word forms is generated on the basis of the transcription. The orthographic forms of the words can then be converted to phonemes by means of computerised grapheme-to-phoneme conversion. The result of this process is a list of citation forms, also called canonical forms or citation-phonemic forms. These forms represent the pronunciation of words when spoken in isolation, and do not cover variations in pronunciation found in running speech. However, this procedure will at least give a standard pronunciation as a starting-point. This is especially relevant if a corpus is to be used by other persons than those belonging to that language community. On the basis of these canonical forms, phonetic transcriptions can be made semi-automatically using large vocabulary speech recognisers.
- 7. If there is no compelling reason otherwise, do not start to transcribe a corpus phonetically, since the time spent on this will never be recovered. If very specific phonetic details are needed, one is advised to look for these on the basis of orthographic and/or phonemic transcriptions.

- 8. It is recommended that transcribers give information about the process of transcribing and about the speech that they have transcribed. Some speakers will be easier to transcribe than other speakers. This will depend on the speech rate, the clarity of articulation, the amount of hesitation, and the number of dialect words used by the speakers. Some information about the difficulty of the transcription is very useful for later queries. The transcribers of the Switchboard (telephone) Corpus were asked to indicate on a scale ranging from 1 to 5 the following characteristics of a conversation: difficulty, topicality, naturalness, echo from B (in listening to A separately, B could hardly be heard (1) or was as nearly as loud as A (5)), echo from A, static on A (no static noise (1) or great deal of it (5)), static on B, background A, and background B.
- 9. In the case of transcriptions at more than one level (e.g. orthographic transcription with some prosodic marks and indications of hesitations etc.), the recommendation is to listen to one level at a time. In everyday life, listeners are accustomed to ignoring hesitations, false starts, and other imperfections, and also do not pay explicit attention to prosody. Transcribers must learn to hear all these events. It seems easiest to listen to the words first and transcribe these, and then to assign the prosodic marks and other annotations.
- 10. For orthographic transcriptions it is not necessary to find experienced transcribers. However, for phonemic and phonetic transcriptions it is necessary to use transcribers who are accustomed to listening to speech in a very precise, analytical way.
- 11. To give some indication about the time necessary to transcribe speech, here are some examples. The time that will be necessary to make an orthographic transcription of spontaneous speech is about ten times the duration of the speech itself. The time necessary for an orthographic transcription of read sentences is about three times the duration of the speech and for an orthographic transcription of read texts it is about five times the duration of the speech.
- 12. Checking of transcription is always necessary. Checking can be done in different ways. An independent transcriber can transcribe the whole or a sample of the corpus. Another possibility is to allow someone else to check the transcription by reading the transcription and listening to the speech. This is less time-consuming. In the case of the latter procedure, it is recommended that the transcription be checked in the opposite order to that used by the first transcriber, since towards the end of the material the first transcriber will be more self-consistent than at the beginning. Inconsistencies may occur in the conventions used (spelling and annotation conventions (brackets, etc.)), as well as in what is heard by the two different persons.
- 13. For the label file format, use any format that can easily be converted to a WAVES label file, for the sake of portability across different systems.
- 14. Any accuracy measure based on inter-transcriber consistency must control for the factors "level of transcription", "segment type", and "task type"

(whether segmentation or labelling).

- 15. If the corpus is confined to one language, and if the labelling is to be alphabetic rather than true IPA symbols, then it is advisable to use a language-specific set of characters. This avoids the notational complexity necessary when all symbols must be kept distinct across all languages, as is needed in the study of general phonetics.
- 16. When transcribing prosodically, the *provisional* recommendation is to use either the ToBI or the IPO system (and the MARSEC system if a purely auditory transcription is being carried out). If the language to be transcribed is not English, and especially if the projected application of the prosodic transcription is in the field of speech technology, then it is probably best to use the IPO system if possible (i.e. if the basic "grammar" of contours has already been researched for that language).

0	в	32	y	64	∞	96	3	
1	a	33	x	65	Ð	97	3	
2	α	34	h	66	þ	98	2	
3	υ	35	ħ	67	φ	99	ſ	
4	Λ	36	$\mathbf{\hat{h}}$	68	ſ	100	5	
5	đ	37	ſj	69	r	101	2	
6	b	38	Ч	70	r	102	ና	
7	þ	39	i	71	I	103	п	
8	6	40	ł	72	ત	104	1	
9	β	41	ι	73	1	105	1	
10	¢	42	Ι	74	R	106	1	
11	ç	43	Ŧ	75	R	107	٦	
12	С	44	ł	76	ş	108	⊥	
13	đ	45	ł	77	ſ	109	т	
14	\mathbf{d}	46	ł	78	L	110	Ч	
15	ø	47	ł	79	σ	111	⊢	
16	ď	48	l	80	t	112	۲	
17	d,	49	Ъ	81	ţſ	113	I	
18	dз	50	λ	82	Ĵ	114		
19	ð	51	χ	83	θ	115	¢	
20	D	52	ŋ	84	ŧ	116	^	
21	ə	53	u	85	уĭ	117	v	
22	$\mathfrak{I}^{\mathrm{b}}$	54	պ	86	υ	118	<	
23	е	55	ր	87	U	119	>	
24	3	56	ŋ	88	Ð	120	0	
25	3	57	η	89	υ	121	0	
26	3^{ι}	58	Ν	90	Μ	122	c	
27	в	59	\odot	91	χ	123	~	
28	g	60	θ	92	λ	124	~	Ì
29	g	61	Э	93	Υ	125	~	1
30	\mathbf{G}	62	ω	94	Z	126	د	1
31	γ	63	ω	95	Z,	127	0	1

Table 5.1: Table of number-to-symbol correspondence for the ${\rm IAT}_{\rm E}{\rm X}$ font wsuipa11.

Part II

Spoken language characterisation

6 Spoken language lexica

6.1 Introduction

6.1.1 Lexica for spoken language systems

Spoken language systems are becoming increasingly versatile, and a central task in developing such a system is the collation of lexical information. Lexical information is required both as a means of characterising properties of words in a spoken language corpus (see Chapter 5) in a lexical database or knowledge base, and for the development of practically all system components. In related areas such as natural language processing (NLP) and computational and theoretical linguistics, the lexicon has come to play an increasingly central role. The lexicon of a spoken language system may be designed for broad or narrow coverage, for specific applications, with a particular kind of organisation, and optimised for a specific strategy of lexical search. Since the construction of a lexicon is a highly labour-intensive and thus also error-prone job, a prime requirement is for formalising lexical representations and automating lexicon development as far as possible, and in re-using lexical resources from existing applications in new developments. The main object of this chapter is to provide a framework for relating such concepts to each other and for the formulation of recommendations for development and use of lexica for spoken language systems.

In this introductory section, some basic concepts connected with the use and structure of lexica in spoken language systems are outlined. In the following sections, specific dimensions of spoken language lexica are discussed in more detail. Particular attention is paid to lexical properties related to *inflectional morphology*, which is far more important for many other languages than it is for English, and other aspects of morphology which are important for the treatment of out-of-vocabulary words. Discussion is restricted to spoken language lexica as system development resources; non-electronic lexica for human use (e.g. pronunciation dictionaries in book form) are not considered. Features common to spoken and written language lexica, such as syntactic and semantic information in lexical entries, are only mentioned in passing; see the report of the EAGLES Working Group on Computational Lexica on these points. The close relation between spoken language lexica and speech corpora results in overlap with the Spoken Language Corpus chapter of this handbook.

The following sections of the chapter are concerned with basic features of spoken language lexica, types of lexical information, lexicon structure, lexical access, and lexical knowledge acquisition for spoken language lexica.

6.1.2 Lexical information as properties of words

At the present time, information about lexica for spoken language systems is relatively hard to come by. One reason for this is that such information is largely contained in specifications of particular proprietary or prototype systems and in technical reports with restricted distribution. With the advent of organisations for coordinating the use of language resources, such as ELRA (the *European Language Resources Association*) and the LDC (the *Linguistic Data Consortium*), access to information on spoken language lexica is becoming more widely available.

Another reason for difficulties in obtaining information about spoken language lexica is that there is not a close relation between concepts and terminology in the speech processing field on the one hand, and concepts and terminology in traditional lexicography on the other. natural language processing and computational linguistics. Components such as Hidden Markov Models for word recognition, stochastic language models for word sequence patterns, grapheme-phoneme tables and rules, word-oriented knowledge bases for semantic interpretation or text construction are all concerned with the the identity and properties of words, lexical access, lexical disambiguation, lexicon architecture and lexical representation, but these relations are not immediately obvious within the specific context of speech technology. Stochastic word models, for instance, would not generally be regarded as a variety of lexicon they evidently do provide corpus-based lexical information about word collocations.

A terminological problem should be noted at the outset: in the spoken language technologies, the term *linguistic* is often used for the representation and processing in sentence, text and dialogue level components, and *acoustic* for word models. With present-day systems, this terminology is misleading. The integration of prosody, for example, requires the interfacing of acoustic techniques at sentence, text and dialogue levels, and linguistic analysis is involved at the word level for the specification of of morphological components in systems developed for highly inflecting languages or for the recognition of out-of-vocabulary words, or for using phonological information in structured Hidden Markov Models (HMMs).

It is useful to distinguish between system lexica and lexical databases. The distinction may, in specific cases, be blurred, and the unity of the two concepts may also be rather loose if the system lexicon is highly modular, or distributed among several system components, or if several different lexical databases are used. However, the distinction is a useful one. The distinction between lexica and lexical databases will be discussed below. Since the kinds of information in both these types of lexical object overlap, the term "spoken language lexicon" will generally be used in this chapter to cover both types. The following overview is necessarily selective.

178 Part II: Spoken language characterisation

6.1.2.1 Types of application for spoken language lexica

Lexica for spoken language are used in a variety of systems, including the following:

- Automatic spelling correctors (spelling is determined to a large extent by phonological considerations).
- Medium and large-vocabulary automatic speech recognition (ASR), as in systems such as SPICOS (cf. Höge et al. 1985; Dreckschmidt 1987; Ney et al. 1988; Thurmair 1986), HEARSAY-II (cf. Lesser et al. 1975; Erman 1977; Erman and Lesser 1980; Erman and Hayes-Roth 1981), SPHINX (cf. Lee et al. 1990), ISADORA (cf. Schukat-Talamazzini 1993), or, for example in automatic dictation machines such as IBM's TANGORA (cf. Averbuch et al. 1986, 1987; Jelinek 1985) and DragonDictate by Dragon Systems (cf. Baker 1975a,b, 1989; Baker et al. 1992).
- Speech synthesis in text-to-speech systems, for example in reading machines, speaking clocks. For further speech synthesis applications, various relevant studies such as Allen et al. (1987), Bailly and Benoît (1992), Bailly (1994), Van Coile (1989), Klatt (1982, 1987), Hertz et al. (1985), Van Hemert et al. (1987) can be consulted (see also Chapter 12).
- Interactive dialogue systems, with speech front ends to databases and enquiry systems and synthesised responses (see for instance Brietzmann et al. 1983; Niemann et al. 1985, 1992; Bunt et al. 1985); see also Chapter 13 on interactive dialogue systems.
- Speech-to-speech translation systems as developed in the ATR and VERB-MOBIL projects, which use various speech recognition techniques, including continuous speech recognition, recognition of new words, word spotting in continuous speech. For speech translation systems see for instance Rayner et al. (1993) and Woszczyna et al. (1993).
- Lexica and encyclopaedias on CD-ROM with multimedia (including acoustic) output.
- Research and development of spoken language processing systems, in the process of which broader based lexica for written language, coupled with tools such as grapheme-phoneme converters, may be used as sources of information.

6.1.2.2 Spoken language lexical databases as a general resource

Spoken language lexica may be components of systems such as those listed above, or reusable background resources. System lexica are generally only of local interest within institutes, companies or projects. Lexical databases as reusable background resources which are intended to be more generally available raise questions of standardised representation, storage and dissemination. In general, the same principles apply as for Spoken Language Corpora: they are collated, stored and disseminated using a variety of media. In research and development contexts, magnetic media (disk or tape) were preferred until recently; in recent years, local magnetic storage and wider informal dissemination within projects or other relevant communities is conducted via the Internet using standard file transfer protocols, electronic mail and World-Wide Web search and access. Large lexica, and corpora on which large lexica are based, are also stored and disseminated in the form of ISO standard CD-ROMs.

The following brief overview can do no more than list a number of examples of current work on spoken language lexicography. At this stage, no claim to exhaustiveness is made, and no valuation of cited or uncited work is intended.

- A number of general lexica with information relevant to spoken language have already been available on CD-ROM for quite some time, including the Hachette and Robert (9 volume) dictionaries for French, the Oxford English Dictionary, the Duden dictionary for German, and the Franklin Computer Corporation Master 4000 dictionary with acoustic output for 83000 words (cf. Goorfin 1989).
- Several lexica with more restricted circulation have been developed in the context of speech technology research and development. Companies such as IBM, and telecom research and development institutes such as CNET in France have developed large lexica (CNET, for instance, has a 55000 word and 12000 phrase lexicon).
- University and other research institutes have also constructed large lexica; in France, for example, such institutes as ENST in Paris, ICP in Grenoble (cf. Tubach and Bok 1985), Paris (cf. Plenat 1991), for a pronunciation dictionary of abbreviations) and IRIT in Toulouse (the BDLEX project) have worked on large spoken language lexica. The BDLEX-1 lexicon coordinated by IRIT (cf. Pérennou and De Calmès 1987) contains 23000 entries, and BDLEX-2 (cf. Pérennou et al. 1991, 1992; Pérennou and Tihoni 1992) contains 50000 entries; a set of linguistic software tools permits the construction of a variety of daughter lexica for spelling correction and lemmatisation, and defines a total of 270000 fully inflected forms.
- The Belgian BRULEX psycholinguistic lexicon contains information on uniqueness points (the point in a letter tree where a word form is uniquely identified), lexical fields, phonological patterns and mean digram frequencies for 36000 words (cf. Content et al. 1990).
- In the United Kingdom, the Alvey project resulted in many tools and lexical materials (cf. Boguraev et al. 1988).
- In the Netherlands, the Nijmegen lexical database CELEX (cf. Baayen 1991), also available on CD-ROM, contains components with 400000 Dutch forms, 15000 English forms and 51000 German forms, together with an access tool FLEX.
- For German, lexical databases for spoken language lexica have been constructed by companies such as Siemens, Daimler-Benz, IBM and Philips, as well as in university speech technology departments (e.g. Munich, Erlangen, Karlsruhe, Bielefeld), and in the VERBMOBIL project (Gibbon 1995; Gibbon and Ehrlich 1995); these have been made available on the World-Wide Web with interactive form interfaces.

180 Part II: Spoken language characterisation

- Work in computational lexicology and computational phonology has led to the development of structured lexicon concepts for spoken language such as ILEX (cf. Gibbon 1992a; Bleiching 1992) based on the DATR lexical knowledge representation language (cf. Evans and Gazdar 1989, 1990); the DATR language has been applied to word form lexica in the multilingual SUNDIAL project (cf. Andry et al. 1992) by the German partner Daimler-Benz and in the German VERBMOBIL project (cf. Gibbon 1993).
- The European Commission has funded a number of projects, particularly within the ESPRIT programme, in which questions of multilingual spoken language system lexica have been addressed, albeit relatively indirectly (POLYGLOT, SUNDIAL, SAM, SAM-A), as well as lexicography projects such as MULTILEX in the ESPRIT programme (cf. Heyer et al. 1991), GENELEX in the EUREKA programme (cf. Nossin 1991) and ACQUILEX, which concentrate on multi-functional written language lexica, though extension of the results to spoken language information has been provided for by the adoption of general sign-based lexicon architectures (see the results of the EAGLES Working Group on Computational Lexica).

6.1.2.3 Lexica in selected spoken language systems

The range of existing spoken language systems is large, so that only a small selection can be outlined, concentrating on well-known older or established systems whose lexicon requirements are representative of different approaches and convey the flavour of basic lexical problems and their treatment. The situation is currently undergoing a process of rapid development. Small vocabulary systems are also excluded, as their strong points are evidently not in the area of the lexicon. The concepts referred to in the descriptions are discussed in the relevant sections below. Reference should also be made to Chapters 5 and 7.

HARPY was a large-vocabulary (1011 words) continuous speech recognition system. It was developed at Carnegie Mellon University. HARPY was the best performing speech recognition system developed under the five-year ARPA project launched in 1971. HARPY makes use of various knowledge sources, including a highly constrained grammar (a finite state grammar in BNF [Backus Naur Form] notation) and lexical knowledge in the form of a pronunciation dictionary that contains alternative pronunciations of each word. Initial attempts to derive within-word phonological variations with a set of phonological rules operating on a baseform failed. A set of juncture rules describes inter-word phonological phenomena such as /p/ deletion at /pm/ junctures: /helpmi/ - /helmi/. The spectral characteristics of allophones of a given phoneme, including their empirically determined durations, are stored in phone templates. The HARPY system compiles all knowledge into a unified directed graph representation, a transition network of 15,000 states (the so-called blackboard model). Each state in the network corresponds to a spectral template. The spectra of the observed segments are compared with the spectral templates in the network. The system determines which sequence of spectra, that is, which path through the network, provides the best match with the acoustic input spectral sequence. (Cf. Klatt (1977); see also Lowerre and Reddy (1980)).

- HEARSAY-II also used the blackboard principle (see HARPY), where knowledge sources contribute to the recognition process via a global data base. In the recognition process, an utterance is segmented into categories of manner-of-articulation features, e.g. a stop-vowel-stop pattern. All words with a syllable structure corresponding to that of the input are proposed as hypotheses. However, words can also be hypothesised top-down by the syntactic component. So misses by the lexical hypothesiser, which are very likely, can be made up for by the syntactic predictor. The lexicon for word verification has the same structure as HARPY; It is defined in terms of spectral patterns. (Cf. Klatt (1977), see also Erman (1977) and Erman and Lesser (1980)).
- SPHINX is a large-vocabulary continuous speech recognition system for speakerindependent application. It was evaluated on the DARPA naval resource management task. The baseline SPHINX system works with Hidden Markov Models (HMMs) where each HMM represents a phone. The total of phones is 45. The phone models are concatenated to create word models, which in turn serve to create sentence models. The phonetic spelling of a word was adopted from the ANGEL System (cf. Rudnicky et al. 1987). The SPHINX baseline system has been improved by introducing multiple codebooks and adding information to the lexical-phonological component:
 - The most likely pronunciation was substituted for the baseform pronunciation of a lexical item in the pronunciation dictionary, retaining the assumption that each lexical item has only one pronunciation.
 - Different models were created for phones that have typically more than one realisation such as released and unreleased /d/ at the beginning of /dIdmaI/ and before /m/, respectively.
 - Two subword units were introduced: function word dependent phone models and generalised triphone models. Since function words are typically unstressed, phones in function words are very often deleted or reduced, do not serve as proper models for recognition tasks, and account for almost 50 % of the errors.

The SPHINX system works with grammars of different perplexity (average branching factor; see Chapter 7); the grammars are of a type which can, in principle, be regarded as a specialised tabular, network-like or treestructured lexicon with probabilistic word-class information:

- A null grammar with a perplexity of 997 (i.e. a vocabulary of 997 words was used); in a null grammar any word can succeed a given word.
- A word-pair grammar with a perplexity of 60; word-pair grammars are lists of words that can follow a given word.

- 182 Part II: Spoken language characterisation
 - A bigram grammar with a perplexity of 20; this is a word-pair grammar equipped with word-category transition probabilities.

In word recognition tests, the best results were obtained with the bigram grammar, the most restrictive kind of the grammars mentioned above (96% accuracy compared with 71% for null grammars).

The Sphinx system has various levels of representation for linguistic units:

- phone models (generalised triphones and extra models for function words),
- word models (stored in the pronunciation dictionary with one representation for each word),
- sentence models (for final confirmation).

(Cf. Lee et al. (1990); see also Alleva et al. (1992)).

- EVAR ("Erkennen Verstehen Antworten Rückfragen", "Recognition Understanding – Answering – Clarification") is a large-vocabulary continuous speech recognition and dialogue system. It is designed to understand standard German sentences and to react either in form of an answer or a question referring back to what has been said, within the specific discourse domain of enquiries concerning Intercity timetables. The EVAR lexicon has the following properties:
 - The lexicon includes not only sublanguage-specific words but also many words of the general vocabulary a dialogue of this kind.
 - The lexicon contains fully inflected word forms.
 - The baseforms, so-called *Normalformen*, e.g. infinitive for verbs, nominative singular for nouns, contain information relevant for all grammatical forms, thus reducing redundancy in the lexicon.
 - The lexicon contains phonological, syntactic, semantic, and pragmatic information.
 - Since the system modules need access only to special lexical knowledge (the *articulation module* makes use of phonological information, while the module in charge of generating the surface structure of an answer also needs syntactic information), access of individual modules to the lexicon is restricted. Preprocessors extract the subset of information relevant for each module.
 - The lexical unit in the EVAR lexicon is the graphemic word (graphematisches Wort); so-called phonetic words (standard pronunciation) and so-called grammatical words (syntactic categories plus meanings) are assigned to the graphemic words.
 - Lexical units are described in attribute-value notation. For example, the attribute WORT takes a graphemic word as its value.
 - Graphemic words again have the attributes AUSSPRACHE (pronunciation) and SYNTAX-TEIL (syntactic part) for which values are defined in the form of a Duden standard pronunciation and morpho-syntactic properties such as the attribute-value pair WORTART-Verb. Numbers keep track of the various entries for different meanings or syntactic variants (e.g. reflexive – non-reflexive),

etc. of a lexical item.

- In the baseform entries, information on stem, pronunciation of the stem (in ASCII symbols that replace the standard IPA notation), and the inflection pattern is given under SYNTAX-TEIL.
- Semantic information includes specifications of semantic features and valence properties as well as selectional restrictions. Fillmore's system of deep structure cases as suggested in Fillmore (1968) has been expanded to 28 cases.

A lexicon administration system has been developed which uses tools for extracting words according to specified criteria, such as "Look for nouns that express a location" or "Look for prepositions that express a direction".(Cf. Ehrlich 1986; Brietzmann et al. 1983; Niemann et al. 1985, 1992)

VERBMOBIL The VERBMOBIL speech-to-speech translation prototype uses lexical information in a wide variety of ways, and much effort went into the creation of standardised orthographic transcriptions, pronouncing dictionaries with integrated prosodic and morphological information, as well as lexica for syntactic, semantic, pragmatic and transfer (translation) information. The system lexicon is distributed between a large number of modules concerned with recognition, parsing, semantic construction and evaluation, transfer, language generation and synthesis, related by "VERBMOBIL interface terms", i.e. standardised lexical information vectors. The VERBMOBIL lexical database was made available to the consortium by means of an interactive World-Wide Web form interface together with a concordance for linguistic analysis, and additional special interactive tools for investigating the phonetic similarities which cause false analyses and misunderstandings and can be used to trigger clarification dialogues (see Chapter 13). The core of the VERBMOBIL lexical database is a knowledge base of 10000 lexical stems, and a DATR/Prolog inference machine which generates 50000 fully inflected forms and 300000 mappings between inflected forms and morphological categories ((cf. Bleiching et al. 1996)).

6.1.3 Recommendations on resources

The following recommendations should be seen in conjunction with recommendations made after the following more specialised sections of this chapter.

- 1. In preparation for decisions on the use of resources, distinguish between *lexical database* and the *system lexicon*.
- 2. Identify the types of information required for the lexical database and the system lexicon.
- 3. Consider the relevant lexical database models and system lexicon architectures.
- 4. Develop a systematic concept for the tools required in producing and accessing a lexicon or a lexical database, prior to construction of such a

184 Part II: Spoken language characterisation

database.

- 5. For information on resources in specific areas of spoken language systems, consult the other chapters and appendices in this handbook.
- 6. For information on resources which relate to written language, consult the other EAGLES working groups.
- 7. For general information on resources consult the organisations ELRA (*European Language Resources Association*) and LDC (*Linguistic Data Consortium*).

6.2 What is a spoken language lexicon?

6.2.1 Basic features of a spoken language lexicon

A spoken language lexicon may be a component in a system, a system lexicon, or a background resource for wider use, a *lexical database*, in each case containing information about the pronunciation, the spelling, the syntactic usage, the meaning and specific pragmatic properties of words; lexica containing subsets of this information may also be referred to as spoken language lexica, though the simpler cases are often simply referred to as wordlists. Where there is little danger of confusion, the term spoken language lexicon will be used to refer indifferently to either a spoken language system lexicon or a spoken language lexical database. A lexical database may be general purpose, or orientated towards specific tasks such as speech recognition or speech synthesis, and restricted to a specific scenario. For system development and evaluation it is generally critical to define an agreed word-list with a well-defined notion of word (e.g. a fully inflected word form), and an associated complete and consistent pronunciation dictionary for grapheme-phoneme conversion and language model construction (see Chapter 7).

A spoken language lexicon is defined as a list of representations of lexical entries consisting of spoken word forms paired with their other lexical properties such as spelling, pronunciation, part of speech (POS), meaning and usage information, in such a way as to optimise lookup of any or all of these properties. This definition covers a wide range of specific types of spoken language lexicon, . At the one end of the spectrum are lists in which orthography provides a more or less indirect representation of a spoken word form pronunciation augmented by tabular pronunciation dictionaries and conversion rules. At the other end are declarative knowledge bases with attribute-value matrix representation formalisms and inheritance hierarchies with associated inference machines, by means of which details of lexical information are inferred from specific premises (entries) about individual lexical items and general premises (rules) about the structure of lexical items. Between these extremes are optimised representations such as those discussed in Chapter 7, and other application directed special lexicon types based, for instance, on the different requirements for pronunciation

tables for speech recognisers and for speech synthesisers.

Both in speech recognition and in speech synthesis, the different kinds of spoken language lexicon are generally orientated towards the *forms* of words rather than towards their *distribution* in larger text or utterance units, or their *meaning* and *use* in context. Furthermore, where possible closed sets of fully inflected words which are actually attested in corpora are preferred to the construction of words on morphological principles, though rule-based word construction is increasing in importance in projects concerned with highly inflecting languages or aimed at the recognition of spontaneous continuous speech in which out-of-vocabulary words or ad hoc coinages (nonce forms) are encountered. In addition to out-of-vocabulary words, systematic noise events may also require inventarisation in a lexical database.

6.2.2 Lexical databases and system lexica for spoken language

The distinction between lexical databases and system lexica is a useful one, though in practice more complex distinctions are required. The main characteristics of the two kinds of lexical object are outlined below.

- LEXICAL DATABASE: A spoken language lexical database is often a set of loosely related simpler databases (e.g. pronunciation table, index into a signal annotation file database, stochastic word model, linguistic lexical database with syntactic and semantic information).
 - Purpose:
 - Resource for system development (training, evaluation; construction of stochastic language models).
 - Definition of vocabulary coverage.
 - Basis for vocabulary consistency maintenance.
 - Reference point for integrating different kinds of lexical information.
 - Source of information for investigation of vocabulary structure.
 - Structure:
 - Generally fixed record structures, with fields for different types of lexical information, and strings as values in fields.
 - Often identification of *lexical key (lexical identifier)* with *orthographic word form*. A problem with orthographic keys, particularly with large vocabularies: is the existence of homographs, i.e. lexical items with the same spelling but different pronunciation (heterophonous homographs) and/or meaning, a potential source of "orthographic noise". Additional serial numbering may be used to distinguish between homographs.
 - Alternative for larger databases with more complex linguistic information: Unique identification of word as a more abstract

unit with a formal identifier and specific properties including orthography, pronunciation, syntax (POS), semantics, etc. on an equal footing.

- Implementation generally conforming to local laboratory standards as a database of ASCII strings, created and accessed by means of standard UNIX tools and UNIX shell scripts, or C programmes; in more complex environments with a commercial database such as ORACLE; occasionally as knowledge bases in higher-level languages such as Prolog or specialised languages such as DATR.
- Content:
 - Main lookup key (in general an orthographic representation, perhaps supplemented by numbering to distinguish homo-graphs).
 - Database entries may be *fully inflected forms, uninflected stems, or morphemes* (generally *morphs, i.e. the phonemic forms of morphemes*), or all of these; other inventories containing units such as phonemes, diphones or syllables, may be included.
 - Pronunciation (in canonical phonemic representation, perhaps including pronunciation variants.
 - Subword boundaries between units such as syllables, morphs (phonemic forms of affixes, lexical roots), derived stems and constituents of compound words.
 - Syntactic category (part of speech, POS, e.g. Noun, Adjective, Article, Pronoun, Verb, Adverb, Preposition, Conjunction, Interjection) or subcategory (e.g. Proper vs. Common Noun, Intransitive vs. Transitive vs. Ditransitive vs. Prepositional, etc., Verb).
 - Semantic categories (in general scenario-specific, i.e. restricted to a given domain or application).
 - Corpus information: frequency statistics (of varying complexity, up to sophisticated language models, cf. Chapter 7); concordance information (i.e. list of contexts of occurrence for each word, usually generated on demand); signal annotations.
 - Further information: concordance (textual context), links to speech files.
 - Implementation:
 - commercial relational or object-oriented database,
 - UNIX ASCII database core with access by UNIX $\,$
 - script languages, C or C++ programmes,
 - in-house custom databases or knowledge bases.
- SYSTEM LEXICON: Lexical information (i.e. properties of words) referred to during the speech recognition or synthesis process may not be concentrated in one identifiable lexicon in a given system.

- Purpose: Definition of those properties of words required for recognition, parsing and understanding, or for planning, formulation and synthesis.
- Structure: In general separate modules for different properties of words with different functions within the system (which are often not regarded as having anything at all to do with a lexicon)
 - In speech recognition: Modules such as the word recogniser (typically based on *Hidden Markov Model* technology), which identifies word forms, i.e. recognition oriented lexical access keys, often phoneme strings derived from orthographic keys and a pronunciation dictionary, the stochastic language model (which defines statistical properties of words in their immediate contexts as *bigrams*, *trigrams*, etc.), and the linguistic lexicon with syntactic and semantic information, linked to an application-specific database or knowledge base.
 - In speech synthesis: Modules which map orthographic forms (in text-to-speech systems) or conceptual or semantic representations (in concept-to-speech systems) to word structures in terms of morpheme sequences, word prosody (e.g. accentuation), and pronunciation (in terms of phonemes), supplemented by detailed rules for phoneme variants in different contexts and for timing and other relevant parametric information.
- Content: Application specific; subsets of information defined in the lexical database resource, as outlined under "Structure".

6.2.3 Spoken language and written language lexica

Spoken language lexica differ in coverage and content in many respects from lexica for written language, although they also share much information with them. Written language lexica are generally based on a stem, neutral or canonical morphological form (e.g. nominative singular; infinitive), or headword concept, in which generalisations over morphologically related forms may be included. This principle leads to fairly compact representations. Spoken language lexica for speech recognition are generally based on fully inflected word forms, as in dictation systems with about 20000 entries. Depending on the complexity of inflectional morphology in the language concerned, the number of fully inflected word form entries is larger than the number of regularly inflectable entries in a dictionary based on stems or neutral forms by a factor from 2 or 3 to several thousand, depending on the typology of the language concerned. Speech synthesis systems for textto-speech applications do not rely exclusively on extensive lexica, but also use rule-based techniques for generating pronunciation forms and prosody (speech melody) from smaller basic units.
An orthographically oriented lexicon generally includes a canonical phonemic transcription, based on the citation form of a word (the pronunciation of a word in isolation) which can be utilised, for example, in sophisticated tools for automatic spelling correction or "phonetic search" in name databases. However, this is not always adequate for the requirements of speech recognition systems, in which further details are required.

A spoken language lexicon may also contain information about pronunciation variants, and often includes prosodic information about syllable structure, stress, and (in tone and pitch accent languages) about lexical tone and pitch accent, with morphological information about division into stems and affixes. Spoken language lexica are in general much more heavily orientated towards properties of word *forms* than towards the *distributional* and *semantic* properties of words.

It may happen that a canonical morphological form or a canonical pronunciation does not actually occur in a given spoken language corpus; this would be of little consequence for a traditional dictionary, but in a spoken language dictionary it is necessary to adopt one of the following solutions (see also Chapter 7 for a discussion of solutions to the sparse data problem in language modelling):

- 1. Use the canonical phonemic form, but mark it as non-occurring; additionally, incorporate the attested form.
- 2. Adopt an attested form as canonical morphological form (e.g. nouns occurring only in the plural such as French *ténèbres* 'darkness', English *trousers*, German *Leute* 'people').

At a more detailed level, orthography (the division of word forms into standardised units of writing) and phonology (the division of word forms into units of pronunciation) are related in different ways in different languages both to each other and also to the morphology (the division of word forms into units of sense) of the language. The orthographic notion of "syllable" serves, in general, in written language lexica for defining hyphenation at line breaks and certain spelling rules (and may even refer to morphological prefixes and suffixes); for this purpose, morphological information about words is also generally required. In spoken language, however, the phonological notion of "syllable" is quite different; it refers to units of speech which are basic to the definition of the well-formed sound sequences of a language and to the rhythmic structure of speech, and forms the basis for the definition of variant pronunciations of speech sounds. Alphabetic orthography involves a close relation between characters and phonemes; in syllabic orthography (Japanese 'Kana') characters are closely related to phonological syllables; in logographic orthography (Chinese), characters are closely related to simplex words (cf. numerals in European languages: the spelling "7" is pronounced /zi:bən/, /sɛt/, /sɛvən/, and so on).

When complex word forms are put together from combinations of smaller

units, different alternations of orthographic units (letters) often occur at the boundaries of the parts of such words (telephone + y = telephony; lady + s = ladies). Similarly, morphophonemic alternations occur in such positions (wife - /walf/singular vs. wives - /walvz/ plural). Furthermore, additional kinds of lexical unit are required in the lexicon of a spoken language dialogue system: discourse particles, hesitation phenomena, pragmatic idioms, such as greetings, or so-called functional units (sequences of functional words which behave as a phonological unit: n'est-ce pas, /nɛspa/) and clitics (functional words which combine with lexical words to form a functional unit, cf. I'm coming, /aIm kʌmɪŋ/).

6.2.4 Basic lexicographic coverage criteria

Criteria for the coverage of lexica for spoken language processing systems are heavily corpus determined, and differ considerably from criteria for coverage of lexica for traditional computational linguistics and some areas of natural language processing. In theoretical computational linguistics, interests are determined by systematic fragments of natural languages which reveal interesting problems of representation and processing. In natural language processing, maximally broad coverage is often the goal. In spoken language lexica as currently used in speech technology, lexica are always oriented towards a particular well-defined corpus which has often been specifically constructed for the task in hand. When speech technology and natural language specialists meet, for instance in comprehensive dialogue oriented development projects, these differences of terminology and priorities are a potential source of misunderstanding and disagreement, and joint solutions need to be carefully negotiated.

The main coverage criteria for spoken language lexica may be summarised as follows.

- Completeness (all word types in the training corpus and test corpora).
- Minimality (only word types in the training and test corpora).
- Consistency (with respect to the training and test corpora and other related data types).
- Generality (projection of the word form set on to related forms not in the corpus).
- Informativity (the types of lexical information associated with lexical entries).

The first four criteria define quantitative or extensional coverage, the fifth defines qualitative or intensional coverage of the lexicon.

These criteria pertain to words; if other units, such as idioms, are involved, the criteria apply analogously to these.

The first three extensional criteria are essentials for the current state of speech technology. Conventional expectations in written language processing, i.e. in natural language processing and computational linguistics, are

widely different, and are expressed in the fourth criterion. Clearly the second and fourth criteria clash; the relation to relevant corpora must therefore be carefully flagged in a spoken language lexicon. The degree of extentional coverage (which for a speech recognition system generally has to be 100%) is sometimes expressed in terms of the notions of degree of static coverage (ratio of in a corpus which are contained in a given dictionary to the number of words in the corpus) and the degree of dynamic coverage or saturation (the probability of encountering words which have previously been encountered); the latter value is generally higher than the former (cf. Ferrané et al. 1992). On the basis of corpus statistics for typologically different languages such as English (cf. Averbuch et al. 1987) and French (cf. Mérialdo 1988), two languages which differ widely in their inflectional structure (English with few verbal inflections, French with a rich verbal inflection system), interesting quantitative comparisons can be made (cf. Table 6.1).

Table 6.1: Examples of morphological coverage

	Vocabulary (no. of forms)	Static coverage	Dynamic coverage
English	5000	92.5%	
English	20000	97.6%	
French	20000	94.9%	98.2%
French	200000	97.5%	99.5%

6.2.5 The lexicon in spoken language recognition systems

A spoken language recognition system is generally divided into two components: the recognition component and the search component (see Chapters 7 and 10). In the recognition component, intervals of the speech signal are mapped by probabilistic systems such as Hidden Markov Models, Neural Networks, Dynamic Programming algorithms, Fuzzy Logic knowledge bases, to word hypotheses; the resulting mapping is organised as a word lattice or word graph, i.e. a set of word hypotheses, each assigned in principle to a temporal interval in the speech signal. The term *word* is used here in the sense of "lexical lookup key". The keys are traditionally represented by orthography, but would be better represented in a spoken language system by phonemic transcriptions. in order to avoid orthographic noise due to heterophonous homographs. The search component enhances the information from the speech signal with top-down information from a language model in order to narrow down the lexical search space. In spoken language recognition system development, a corpus based lexicon of orthographically transcribed forms is used as the basis for a pronunciation lexicon (pronunciation dictionary); the lexicon is often supplemented by rules for generating pronunciation variants due to informal speech styles (phonostylistics) or speaker and dialect variants. The pronunciation lexicon is required in order to tune the recognition system to a specific corpus by statistical training: frequencies of distribution of words in a corpus are interpreted as the prior (*a priori*) probabilities of words in a given context. These prior probabilities may be based on the absolute frequencies of words, or on their frequencies relative to a given context, e.g. *digram* (*bigram*) frequencies.

The functionality of spoken language lexica may be summarised in the following terms.

- Off-line functions
 - Fully inflected word form list construction
 - Pronunciation table (lexicon) construction
 - Synthesiser development
 - Recogniser (forced alignment, stochastic training)
 - Orthographic transcription (Transliteration) checking
 - Integration of word and sentence prosody
 - Integration of morphology for
 - describing new words
 - sparse data training with stems and word classes
 - Frequency table construction
 - Word distribution frequency tables
 - Language models
 - Coverage definition for inter-project coordination
- On-line functions
 - Search for word forms for inclusion in word lattice
 - orthographic (conventional technology)
 - phonological (new architectures)
 - Definition of criteria for lexicon architecture and lookup

6.2.6 Recommendations on defining spoken language lexica

- 1. Define the anticipated functions of the off-line lexical database in spoken language system development, and the on-line system lexicon components, bearing in mind the differences between written text and speech, the intended application, and possible economies to be gained in using or creating reusable lexical resources.
- 2. Specify the extensional (quantitative) coverage (size) of the lexical database and of the system lexicon with reference to the available system components and in terms of
- 3. Specify the intensional (qualitative) coverage (content) of the lexical database and of the system lexicon components with reference to the application domain and in terms of standard types of lexical information.

- 192 Part II: Spoken language characterisation
- 6.3 Types of lexical information in spoken language lexica
- 6.3.1 Lexicon models and lexical representation

A given system lexicon or lexical database is based on a *lexical information model* or a *data model*; often the model is intuitively constructed, or based on notions taken from traditional school grammar, but scientifically motivated models are becoming available. A model of lexical information will make at least the following distinctions:

- Lexical objects: The basic objects (such as words) described in a lexicon. It is becoming customary in lexicography and computational linguistics to refer to the *lexical sign*, i.e. an object associated with attributes denoting orthogonal kinds of lexical information. A second kind of lexical object is the *lexical sign class* or *archi-sign* in which similar lexical objects are grouped together, each characterised by subsets of the lexical information required to characterise specific lexical signs. These class-based generalisations may be organised in terms of *implication rules (redundancy rules)*, subsumption lattices, *type hierarchies*, or *default inheritance hierarchies*.
 Lexical information: In a theoretically well-founded lexicon which satisfies formal criteria of consistency and coverage criteria such as empirical completeness and soundness, types of lexical information are *orthogonal*, i.e. of different types which complement each other. These orthogonal types of
- lexical information are often labelled with *attribute* names, and the items of information regarded as the *values* of these attributes. Values may be complex, expressed as nested attribute-value structures. The types include orthography, pronunciation, syntactic distributional properties, meaning, and pragmatic properties of use in context (e.g. speech act type, stylistic level). See also the results of the EAGLES Working Group on Formalisms.

Lexicon models for lexical databases and system lexicons are part of the overall conceptual framework required for lexicon development. Modern approaches to lexicon development provide suitable lexical representation languages for formulating and integrating the different kinds of lexical information specified in a lexicon model and assigning them to lexical objects, and implementations for these representation languages (cf. Andry et al. 1992). In recent work, the following useful distinctions are sometimes made:

- Lexicon formalism: A specially designed logic programming language such as DATR, or an algebraic formalism such as attribute-value matrices, or appropriate definitions in high level languages such as LISP or Prolog, with compiler concepts for translating these languages into conventional languages for efficient processing. Imperative languages such as C are sometimes used directly to represent smaller lexicons, or where speed of access is at a premium, but this is not a generally recommended practice.
- Lexicon theory: A coherent and consistent set of expressions formulated in a well-defined formalism and interpreted with respect to a lexicon model.

- General lexicon theory: A general theory of lexical objects and information, for instance a theory of lexical signs and their representation.
- Specific lexicon theory: A given lexicon formulated in a lexicon formalism on the basis of a lexicon model.
- Lexicon model: Specification of the domain denoted by a lexicon theory, conceptually independent of the theory itself (cf. the notion of a *data model* for a database). A different definition is also common: the general structure of the objects and attribute-value structures in a formal lexicon. A lexicon model specifies the following kinds of information:
 - Types of lexical object and structure of lexical entries.
 - Types of lexical information associated with lexical objects in lexical entries.
 - Relations between lexical objects and structure of the lexicon as a whole *lexicon architecture*.
- Linguistic framework: In recent large projects such as VERBMOBIL, general linguistic frameworks such as HPSG (Head-Driven Phrase Structure Grammar) have been used.

The aspects of representation and architecture will be dealt with in a later section. The following subsections are concerned with the main kinds of lexical information required for spoken language lexical entries.

6.3.2 A simple sign model for lexical properties

Lexical information is often regarded as a heterogeneous collection of idiosyncratic information about lexical items. An assumption such as this makes it hard to discuss lexical information systematically and, moreover, from the point of view of contemporary lexicography, it is wrong. For this reason, a simple unifying informal model of lexical signs, related to a view which is current in computational linguistics and computational lexicography, is used for the purpose of further discussion.

In general terms, a sign is a unit of communication with identifiable form and meaning. Lexical signs have specific ranks, such as word or phrase (for phrasal idioms), and include: words, phrasal idioms and other items such as dialogue control particles (er, uhm, aha etc.). It may also be argued that even smaller units such as morphemes also have sign structure. Lexical signs thus range, in principle, over fully inflected word forms, morphs (roots, affixes), stems (roots or stems to which affixation has applied), lemmas (or lemmata), often represented by an orthographic form, and phrasal items (idioms).

Lexical signs are characterised by the following four basic types of information:

- 194 Part II: Spoken language characterisation
- 1. Surface properties: *orthographic* and *phonological* representation; for pronunciation, several different levels of transcription are possible (morphophonemic, phonemic).
- 2. Semantic properties: *semantic* and *pragmatic* representation.
- 3. Distributional properties: syntactic category and subcategory (e.g. Verb, Transitive Verb).
- 4. Compositional properties: *head* and *modifier* (*complement* or *specifier*) constituents; word formation is recursive:
 - [[[[mouse] [trap]] [[repair] [shop]]] [owner]]

The first two types are referred to as *interpretative* properties, since they interpret the basic sign representation in terms of the real world of phonetics (or writing) and the real world of meaning, while the second two types are referred to as *structural* (or *syntactic*, in a general sense of the term) properties. Complex signs are constructed compositionally from their constituent signs and derive their properties compositionally from these. Non-lexical signs include, for example, freely invented compound words, such as the example given above, or almost any sentence in this book.

The following sections will be devoted to the four main types of lexical information, referring to them as *surface*, *content*, *grammatical* and *morphological* information, respectively.

In the examples given below, a basic computer-readable attribute-value syntax is used, based on the kind of spoken language lexical representation in DATR used by Andry et al. (1992). The name of the lexical sign (which is not necessarily its orthography) is written with an initial upper case letter and followed by a colon, attribute names can be either word-like atoms or sequences of atoms (in the latter case, permitting an indirect representation of more complex attribute structures); they are enclosed in corner brackets and separated from their values by an equality sign, and the lexical sign is terminated by a period. The SAMPA notation used below is defined in Appendix B; see also Chapter 5.

```
Table:
```

```
<surface orthography> = table
<surface phonetics sampa> = teIbl
<semantics> = artefactual horizontal surface
<distribution> = noun common countable
<composition> = simplex z_plural.
```

In the case of complex signs, the meaning of the sign is a function of the meanings of its parts and the pronunciation of the sign is a function of the pronunciations of its parts. These functions may be partly idiosyncratic with lexical signs; this is shown in the pronunciation and meaning of words like English "dustman":

```
Dustman:

<surface orthography> = dustman

<surface phonetics sampa> = dVsm@n

<semantics> = 'municipal garbage collector'.
```

The *pronunciation* and *meaning* of this complex lexical sign are not in all respects a general compositional function of its parts, for example the pronunciation of *dustman* is not /dʌstmæn/ but /dʌsmən/, nor is a dustman necessarily only concerned with dust:

In contrast, the *spelling* and the *distribution* of the complex sign are perfectly regular functions of the spellings of the parts and the distribution of the head (i.e. Man) of the sign, respectively.

In perfectly regular cases, there would therefore be no necessity to include complex words in the lexicon. Such cases are practically non-existent, however, since complex words are in general partially idiosyncratic; in a comprehensive spoken language lexicon, both complex words and their parts therefore need to be included. For most current practical purposes, in which *potential words* (unknown words or ad hoc word formations) do not need to be treated in addition to *actual words* (those contained in a lexicon), complex words can be listed in full as unanalysed forms.

Modern computational lexicographic practice attempts to reduce the redundancy in a lexicon as far as possible: fully regular information in compounds can be *inherited* from the parts of the compounds, while idiosyncratic information is specified locally. In a case like this, a lexical class is specified for defining the structure of compounds, and "inheritance pointers" are included. The result is a *hierarchical lexicon structure*, in which macro-like cross-references are made to other lexical signs (analogous to cross-references in conventional dictionaries), but also to whole classes of lexical signs (*archisigns*).

6.3.3 Lexical units

6.3.3.1 Kinds of lexical unit

Intuitively, the prototypic lexical unit is a word. This definition has a number of catches to it, however, because the notion of word is not as

simple as it seems, and because lexical phrases (idioms) also exist. The intuitive notion of *word* has "fuzzy edges", as in the following cases:

- 1. Words may contain other words (e.g. compound words such as *database*, *Sprachtechnologie*).
- 2. Words have different status in respect of their phonetic realisations and their meaning; compare the difference between function words, e.g. to, for with reduced pronunciations and structural meanings, and content words, e.g. word, spell, which refer to real world objects, properties, event types, abstract concepts.
- 3. Words may be merged with other words in informal speech (*cliticisation*). Examples of clitics are English 's in he's - /hi:z/, French l' in il l'a vu - /il la vy:/, German 'm in auf 'm Tisch - /aufm tIf/.
- Particular types of word formation such as spelling and acronym formation may require special attention: ecu - /i:k''ju:/, /i:si:''ju:/.
- 5. Words may be *inflected word forms*, making *sound* (singular) and *sounds* (plural) into different words. On the other hand, words may be regarded as a class of inflectionally related forms (a *paradigm*), i.e. *sound* and *sounds* then belong to the same word, which may be characterised by a canonical inflected form (e.g. nominative singular), or by the stem shared by the forms and identified by linguistic analysis, or by a number or other abstract label. In speech technology, the *inflected word form* is the standard definition. In standard dictionaries, the *paradigm* definition of word is used, represented by a *headword* or *lemma*, generally the canonical inflectional form such as nominative singular, in orthographic representation.
- 6. Lexical units may need to be larger than the word (e.g. phrasal idioms).
- 7. Lexical units may need to be smaller than the word: *Semantically oriented* morphological word subunits (word constituents) include
 - word stems minus inflections; indivisible word stems are *lexical* morphemes);
 - constituent words words formed by compounding (composition);
 - constituent prefixes, stems and suffixes in words formed by derivation.

Pronunciation oriented phonological word subunits include syllables and their parts; phonological subunits do not necessarily correspond closely with morphological subunits.

- 8. Linguistic textbooks distinguish between several different views of words as lexical units, depending on which kind of lexical sign information is regarded as primary:
 - The *phonological word* (based on its conformity to the phonotactic structure of a language).
 - The *prosodic word*, based on its conformity to the accentuation and the rhythm patterning of the language.
 - The *orthographic word* (for instance, as delimited by spaces or punctuation marks).
 - The morphological word (based on the indivisibility and fixed internal structure of words).

- The syntactic word (based on its distribution in sentences).
- 9. The lexical word as a *type*, as opposed to an *occurrence* of the type in larger units, and a *token* of the type in a corpus of speech or writing.

The central meaning for the purpose of spoken language lexica will be taken to be the *morphological word*.

Lexical units (entries, items) are assigned sets of properties; these identify the lexical units as signs, and determine the organisation of the lexicon. In practical contexts, the choice of lexical unit and the definition of priorities among its properties may be important for procedural reasons, i.e. in determining ways in which a lexicon may be most easily accessed: through orthography, pronunciation, meaning, syntactic properties, or via its morphological properties (stem, inflection). The application-driven decision on the kind of lexical unit which is most suitable for a given purpose is a nontrivial one. However, for many practical purposes fairly straightforward guidelines can be given:

- The form of a lexical item, in particular its orthography, is often used as the main identifying property for accessing the lexicon.
- However, access on phonetic grounds, via the phonological form, is evidently the optimal procedure for speech recognition, and access on conceptual semantic or syntactic grounds is evidently the optimal procedure for speech synthesis.
- The use of orthography as an intermediate stage in speech recognition is a useful and widespread heuristic which generally does not introduce significant numbers of artefacts into the mapping from speech signals to lexical items, but is not recommended for complex systems with large vocabularies, except as a means of visualisation in user interfaces.
- For text-to-speech applications orthography is likely to be the optimal lexical access key.

6.3.3.2 Fully inflected form lexica

It has already been noted that fully inflected form lexica and lexical databases are fairly standard for speech recognition. Where a small closed vocabulary is used, and new, unknown or ad hoc word formations are not required (as with most current applications in speech synthesis and recognition), fully inflected word forms are listed. This procedure is most convenient in languages with very small inflectional paradigms; for languages of the agglutinative type, in which large numbers of inflectional endings are concatenated, the procedure rapidly becomes intractable. In other applications, too, such as speech synthesis, it may be more tractable to generate fully inflected word forms from stems and endings.

An example of a language with few inflections is English, where (except for a few pronouns) only nouns and verbs are inflected, and even here three forms exist for nouns (uninflected, genitive and plural) and four for verbs

(uninflected, third person singular present, past, and present participle; irregular verbs in addition have a different past participle form – the verb $to \ be$ is, as always, an extreme case). English is therefore not a good example for illustrating inflectional morphology (in other areas of morphology, i.e. in word formation, languages appear to be equally complex).

French is much more complex, with inflections on adjectives, and large verb paradigms; note that orthographic inflection in French has more inflectional endings than are distinguished in phonological inflection.

German also has complex inflectional morphology, with significantly more endings on all articles, pronouns, nouns, adjectives and verbs, increasing the size of the vocabulary over the size of a stem-oriented lexicon by a factor of about 4.

In extremely highly inflecting languages such as Finnish, the number of endings and the length of sequences of endings multiply out to increase the vocabulary by a factor of over 1000. Special morphological techniques have been developed (e.g. two-level morphology) to permit efficient calculation of inflected forms and to avoid a finite but unmanageable explosion of lexicon size for highly inflecting languages (cf. Koskenniemi 1983; Karttunen 1983). These techniques have so far not been applied to any significant extent in speech technology (but cf. Althoff et al. 1996).

The figures cited refer only to the sets of forms. When the *form-function* mapping, i.e. the association of a given inflected form with a morphosyntactic category, is considered, the figures become much worse. A single inflected adjective form such as *guten* in German has 44 possible interpretations which are relevant for morphosyntactic agreement contexts (cf. Gibbon 1995), with 13 *feminine* readings, 17 masculine readings, and 14 neuter readings, depending on different cases (nominative, accusative, genitive and dative) and different determiner (article) categories (strong, weak and mixed). It is possible to reduce the size of these sets by means of default-logic abbreviations in a lexical database, but for efficient processing, they ultimately need to be multiplied out. Similar considerations apply to other word categories, and to other highly inflecting languages.

Complex inflectional properties in many languages other than English imply that, for these languages, large vocabulary systems with complex grammatical constructions require prohibitively large fully inflected form inventories. Although the sets of mappings involved can be very large, the inflectional systems of languages define a finite number of variants for each stem, and therefore it may make sense in complex applications in speech recognition to define a rule-based "virtual lexical database" or "virtual lexicon" which constructs or analyses each fully inflected word form on demand using a *stem* or *morph lexicon* with a *morphological rule component* (Althoff et al. 1996; Bleiching et al. 1996; Geutner 1995).

6.3.3.3 Stem and morph lexica

Lexica based on the morphological parts of words, coupled with lexical rules for defining the composition of words from these parts, are not widely used in current speech recognition practice. They are useful, however, in expanding lexica of attested forms to include all fully inflected forms, for instance for word generation and speech synthesis, and in tools which verify the consistency of corpus transcriptions and lexica.

Terminology in this area is somewhat variable. In the most general usage, a *stem* is any uninflected item, whether morphologically simple or complex. However, intermediate stages in word formation by affixation, and in the inflection of highly inflected languages, are also called stems. The smallest stem is a *phonological lexical morph* or an *orthographic lexical morph*, i.e. the phonological or orthographic realisation of a *lexical morpheme*. Since stems may vary in different inflectional contexts, as affixes do, it is necessary to include information about the morphophonological (and morphographemic) alternations of such morphemes:

Knife:

<surface phonology singular> = naIf
<surface phonology plural> = naIv + z
<surface orthography singular> = knife
<surface orthography plural> = knive + s.

The use of morphological decomposition of the kind illustrated here has been demonstrated to bring some advantages in medium size vocabulary speech recognition in German (cf. Geutner 1995); for languages like English, with a low incidence of inflections, the advantage is minimal.

In a stem lexicon, the basic lexical key or *lemma* is the stem, which is represented in some kind of normalised notation. The most common kind of normalised or canonical notation has the following two properties:

- 1. Canonical inflected form: With morphologically inflected items, a "normal form" such as the *infinitive* for verbs or the *nominative singular* for nouns is used.
- 2. Canonical orthography: A standardised orthographic representation of the canonical inflected form is used.

For specific purposes, in which lexical entries need to be accessed on the basis of a specific property, indexing based, for instance, on the canonical phonemic representation, either of a fully inflected form or of the canonical inflected form, or even of the stem itself, may be required; for stochastic language models, for example, a tree-coded representation may be the optimal representation (see Chapter 7). Phonemic representation is dealt with in more detail below.

6.3.3.4 The notion of "lexical lemma"

As in the *knife* example, one particular form, for instance orthographic, of an entry is often used as a headword or lemma. From a technical lexicographic point of view, this form then has a dual function:

- 1. It names the entry.
- 2. It also represents one of its properties, namely its spelling.

In spoken language lexicography, this distinction is central, and ignoring it may lead to confusion. This applies particularly in the context of spoken language lexicography, where the primary criterion of access by word form is phonological.

When homographs occur (e.g. bank as a financial institution or as the side of a river), an additional consecutive numbering is used, e.g. $bank_1$, $bank_2$, etc.

The concept of an *abstract lemma*, deriving from recent developments in computational linguistics and their application to phonology and prosody, may be used in order to clarify the distinction (cf. Gibbon 1992a): an abstract lemma may have any convenient unique name or number (or indeed be labelled by the spelling of the canonical inflected form, as already noted); all properties have equal status, so that the abstract lemma is neutral with respect to different types of lexical access, through spelling, pronunciation, semantics, etc. The examples of lexical entries given so far are based on the concept of an abstract lemma. The neutrality of the abstract lemma with respect to particular properties and particular directions of lexical access make it suitable as a basic concept for organising flexible lexical databases. A lexicon based on a neutral abstract lemma concept is the basic form of a *declarative lexicon*, in which the structure or the lexicon is not dictated by requirements of specific types of lexical access (characteristics of a pro*cedural lexicon*, but by general logical principles. The distinction between declarative and procedural lexica is a relative one, however, which is taken up in the section on spoken language lexicon architectures. For practical applications, a lexical database will need to be procedurally optimised (=indexed) for fast access.

6.3.4 Lexical properties and lexical relations in spoken language

The complex relations between orthographic, phonological, syntactic and semantic properties of lexical units make a theoretically satisfying definition of "lexical sign" quite elusive. Lexical relations are either *paradigmatic*, and define classes of similar items, or *syntagmatic*, and define complex items in terms of relations between their parts.

Present discussion is restricted to the main paradigmatic relations in traditional terms. The expression of these relations in terms of *semantic features, semantic markers* or *semantic components* is not dealt with explicitly, though it figures implicitly in the attribute-value structures which are referred to in the examples.

The syntagmatic relations (semantic roles; collocational relations; syntactic subcategories, valencies) are more complex. Introductions to linguistics may be consulted on syntagmatic relations in sentences (constituent structures and dependency structures). For further information on semantic properties, reference should be made to standard textbooks such as Lyons (1977) or Cruse (1986). Reference should also be made to the results of the EAGLES Computational Lexica Working Group.

The following systematised versions of traditional definitions express the main paradigmatic relations between lexical signs.

- 1. The main relations of *form* between lexical signs are as follows:
 - HOMONYMY: Two words with the same orthographic and phonological forms, but different syntactic categories and/or meanings are homonyms. Example: *mate* /meit/ 'friend' or 'final state of play in a chess game'.
 - HOMOGRAPHY: Two words with the same orthographic form and different phonological forms are (heterophonic) homographs. Example: row /rəu/ 'horizontal sequence', /rau/ 'noise, quarrel'.
 - HOMOPHONY: Two words with the same phonological form and different orthographic forms are (heterographic) homophones. Example: *meet* /mi:t/ 'encounter' *meat* /mi:t/ 'edible animal tissue'.
 - HETEROGRAPHY: Two orthographic forms of the same word are heterographs. Example: *standardise – standardize /*stændadaIz/.
 - HETEROPHONY: Two phonological forms of the same word are heterophones. Example: *either* /aiðə/ - /i:ðə/ 'disjunction'.
- 2. The main relations of *function* between lexical signs:
 - HYPERONYMY: If the meaning of one word is entailed by the meaning of another, it is a hyperonym of the other (a superordinate term relative to the other). Example: *book* is a hyperonym of *manual* as the meaning of *book* is implied by the meaning of *manual* (in one of its readings).
 - HYPONYMY: The converse of hyperonym. If the meaning of one word entails the meaning of another, it is a hyponym of the other (a subordinate term relative to the other). Example: *manual* is a hyponym of *book* as the meaning of *manual* implies the meaning of *book*.
 - CO-HYPONYMY: Two words are co-hyponyms if and only if there is a word which is a hyperonym of each (in the same reading of this word). Example: manual and novel are co-hyponyms in relation to book.
 - SYNONYMY: Two words are synonyms if and only if they have the same meaning (or at least have one meaning in common), i.e. if the meaning of each entails the meaning of the other. They are partial synonyms if either has additional readings not shared by the other. They are full synonyms if they have no reading which is not shared by the

other. Example: manual and handbook are partial synonyms (manual is also, among other things, a term for a traditional organ keyboard). Full synonyms are rare. By implication, synonyms are also co-hyponyms.

ANTONYMY: Two words are antonyms (a) if they are co-hyponyms with respect to given meanings, and (b) if they differ in meaning in respect of those details of the same meaning which are not shared by their hyperonym. Example: *manual* and *novel* are antonyms. Note that the term is sometimes restricted to binary oppositions, e.g. *dead* – *alive*.

In addition to these lexical relations, there are a number of syntagmatic complexities which hold between different types of information.

- Semantically, recursion in word formation is unrestricted, with left- or right branching, or centre-embedding.
- Morphologically, recursion is restricted to flat, linear concatenation, as in: Spracherkennungsevaluationsmethode –
- Sprach#er+kenn+ung#+s#evalu+ation#+s#method+e, or operationalisation - oper+at+ion+al+is+at+ion, which can be efficiently described and implemented by finite state devices.
- Morphophonological modifications of the basic concatenative structure occur, with superimposed word stress or tone patterns, vowel and consonant modifications, as in *telephone - telephony*, *bring - brought*.
- So-called *bracketing paradoxes* occur because of the different morphological structures determined by semantics and phonology; the most wellknown example is *transformational grammarian*, semantically bracketed as ((*transformational grammar*) *ian*), morphologically bracketed as (*transformational (grammar ian*)).
- Note, too, that morphological (lexical, semantic oriented) bracketing does not necessarily correspond with non-lexical, phonologically motivated syllabic bracketing, as in operation operate - ion - /ɔ. pə. r''ei. ∫ən/.

6.3.5 Recommendations on types of lexical information

- 1. Distinguish between the declarative aspect of *types of lexical information* in a lexicon or a lexical database (e.g. orthography, pronunciation, syntactic category) and the procedural aspects of how to use this information in *lexicon development* and for *lexical access* to a lexicon or a lexical database.
- 2. Specify an explicit lexicon model for the intensional coverage and the structure of lexical entries and the overall structure of the lexicon, both for the lexical database and for the system lexicon.
- 3. Define the basic *lexical entry types* or *lexical objects* (e.g. fully inflected forms, morphological stems, conceptual units), and the notation for lexical keys.
- 4. Distinguish clearly between the status of *orthography* as a property of written language, *canonical phonemic representation* as a property of spoken language, and the *lexical key* as a unique identifier for lexical entries (which

may often be the orthography of a canonical inflected form, supplemented by a serial numbers in the case of homographs).

- 5. Define explicit specifications for the notation of orthography and for canonical phonemic representation; for the latter, the IPA-based SAMPA alphabet is recommended (see the following section, and Appendix B).
- 6. First specify the lexicon model suitable for the intended applications, then select formalisms appropriate to the model, i.e. representation conventions for the lexical database and the system lexicon.
- 7. Do not "programme" the lexical information directly in an existing formalism without first explicitly specifying the lexicon model; if possible, compile the system lexicon or pronunciation table (wordlist) automatically from a lexical database or knowledge base.

6.4 Lexical surface information

6.4.1 Orthographic information

Orthography has been used in several different roles in spoken language lexica, some of which have already been noted:

- 1. Convenient general reference labels for words, due to the high level of awareness of, familiarity with and standardisation of orthography in literate societies.
- 2. Convenient identifying names for lexical entries, for "normal lemma" forms, and for headwords in complex lexicon entries which group related words together.
- 3. Convenient identifying names for word hypotheses in word lattices, as lexical lookup keys.
- 4. Visualisation of word hypotheses in a development system.
- 5. Representation of the orthographic properties of words (the main function).

Each of these functions is distinct and needs to be kept conceptually separate in order to avoid confusion. The functions (1) and (2) are not particularly problematic. Function (3) is traditionally a feature of speech recognition systems for relatively small vocabularies. The larger the vocabulary, however, the greater the danger of introducing unnecessary *orthographic noise*, i.e. intrusive artefacts due to homography (words with identical spelling and different pronunciation); for this reason, in new architectures, phonological (e.g. phonemic or autosegmental) representation in word graphs may be preferred. Function (4) is unproblematic, though similar reservations as with (3) are to be noted. Function (5) is the main function and is obviously essential for written output of any kind; however, it is often confused with both functions (2) and (3). Care with consistent orthography is obviously essential.

Orthography has the advantage of being highly standardised, except for certain regional variants (British and American English; Federal, Swiss, and

Austrian German) and variations in publishers' conventions (e.g. British English *ise/-ize* as in *standardisation/standardization*, capitalisation of adjectives in nominal function in German, as *die anderen / die Anderen*, or variations in hyphenation conventions and the spelling of compound words; variation is found particularly in the treatment of derived and compound word s (e.g. separation and hyphenation) and in the use of typographic devices such as capitalisation). Orthography is given further attention in the section on lexical representation.

A standard orthographic transcription is often used for convenience as a means of representing and accessing words in a spoken language lexicon. This has several reasons:

- 1. Familiarity to all educated speakers of the language.
- 2. High level of standardisation in comparison with theory-influenced phonological transcriptions.
- 3. Sufficient proximity to phonological form, at least in European languages, ensures a reasonably close mapping to pronunciation at the level of whole words (not necessarily in the details of grapheme to phoneme mapping) in small vocabularies in some languages (French and English are notorious exceptions).

Most European languages have highly regulated orthographies, the use of which is associated with social and political rewards and punishments. Official orthographic reforms, which typically generate much controversy among the general public, may necessitate some re-implementation of spelling checkers and grapheme-phoneme converters (cf. the ongoing reform of German orthography).

For use in spoken language lexica, particularly in word lists used for training and testing recognisers, consistency is essential and often additional conventions are required in order to meet the criterion of general computer readability in the case of special letters and diacritics. Although it cannot be regarded as a standard, it is becomming common practice to use the ASCII codings or their IATEX adaptations for specific countries. For example, a standard computer-readable orthography for German has become widely accepted for German speech recognition applications which marks special characters, in particular those with an *Umlaut* diacritic, as shown in Table 6.2.

The results of the EAGLES Working Groups on Text Corpora and Lexica should be consulted on orthographic and other matters pertaining to written texts.

6.4.2 Pronunciation information

Pronunciation information is much more application specific (and indeed theory specific) than orthographic information. Standardly, information about phonemic structure is included in the form of a *phonemic transcrip*-

Standard orthography	ASCII orthography
Äpfel	''Apfel
ändern	"andern
Öl	''01
$\ddot{\mathrm{o}}\mathrm{stlich}$	"ostlich
Überzug	''Uberzug
über	"uber
$\mathrm{hei} \mathfrak{B}$	hei″s

Table 6.2: Computer readable ASCII orthography for German

tion of a standard canonical or citation form pronunciation, i.e. the pronunciation of a word in isolation in a standard variety of the language. Often the phonemic transcription is enhanced by including prosodic information such as the stress position (Dutch, English, German), the type of tonal accent (Swedish), syllable and word boundaries in compound words, and word and phrase boundaries in phrasal idioms. Morphological information (morph boundaries, as well as the boundaries of words and phrases) is relevant to stress patterning, and is sometimes also included.

A particularly thorny question is the inclusion of information about pronunciation variants, of which there are two main types, *rule-governed* allophonic and phonostylistic variants, and idiosyncratic *lexical* variants. The following rules of thumb can be given:

- Pronunciation lexica for synthesis generally require one standard (canonical) pronunciation; however, variants of these with different prosodic contexts may be required.
- Pronunciation lexica for recognition require a distinction to be made between variants of the same word, and variants which are associated with the same spelling but different words (heterophonic homographs).
- Strictly speaking, pronunciation lexica for recognition require only lexical variants to be listed which are idiosyncratic and cannot be predicted by rule (e.g. English *either* /aiðə/ /i:ðə/). Variants which are general and regular (such as the reduction of schwa + liquid or nasal to a syllabic liquid or nasal) can be calculated using pronunciation rules (phonological rules): English *running* /rʌniŋ/ /rʌnin/, German *einem* /ainəm/ /ainm/ /aim/).

Although *phoneme* is a technical term with somewhat different definitions in different theoretical contexts, and although there are technical arguments due to Generative Phonology (cf. Chomsky and Halle 1968) which show that the notion of phoneme leads to inconsistencies, the core of phoneme theory is relatively standard. In linguistics handbooks, the phoneme is

commonly defined as the minimal distinctive (meaning-distinguishing) unit (temporal segment) of sound. In the following fairly standard definition, the distinctiveness criterion is implicit in the concept of a *system*; the concept of a *sound* (= *phone*, *allophone*) covers possible variants of a phoneme (e.g. English aspirated word-initial /p/ as opposed to unaspirated /p/ in the context /sp.../ (Crystal 1985, p. 228):

PHONEME (PHONEMIC(S)) The minimal unit in the sound SYS-TEM of a LANGUAGE ... Sounds are considered to be members of the same phoneme if they are phonetically similar and do not occur in the same ENVIRONMENT.

A fairly complete definition is thus based on distinctiveness, minimality, phonetic similarity and distributional complementarity. Phoneme definitions are differential or relational definitions, illustrated by the notion of minimal difference between two words in minimal pairs such as the items in the set of English words pin-tin-kin-fin-thin-sin-shin-chin-bin-din-ginwin-Lynne-Min-Nin, (in standard SAMPA computer readable phonemic transcription: /pIn - tIn - kIn - fIn - θ In - sIn - SIn - tfIn - bIn - dIn dgn - wIn - IIn - mIn - nIn/) (the last three are names). Phonemes defined in this way are further classified as bundles of phonological distinctive features. Operationally, phonemes are defined by procedures of segmentation and classification (reflected, for example, in the recognition and classification components of automatic speech recognition systems):

- Segmentation is the procedure of isolating minimal distinctive temporal phonetic segments (*phones*).
- Classification is the procedure of classifying phones as *allophones* (phonetic *alternants* of the same *phoneme*, on the grounds of distinctiveness, minimality, phonetic similarity and complementary distribution (i.e. their occurrence in complementary contexts as contextual variants of that phoneme).

In contrast to orthographic representations, which for social and cultural reasons, are highly standardised common knowledge, lexical representations of pronunciation are theory and application specific. The most widely used representations in pronouncing dictionaries for human use, such as in foreign language teaching, and in spoken language systems, are *phonemic transcriptions*.

Phonemic descriptions are available for several hundred languages, and phonemic transcriptions based on these are suitable for constructing roman orthographies for languages which have orthographies based on different principles (e.g. syllabic or logographic) or no orthography at all. For a given language, phonemic descriptions differ peripherally (for instance, it is controversial whether diphthongs and affricates are to be analysed as one phoneme or two?). Phonemes are in general the units of choice for practical phonological transcriptions in spoken language system lexica. Other, more specialised types of representation such as the feature matrix representations required by all modern phonological descriptions, and autosegmental lattice representations, or metrical tree graph and histogram representations (cf. Goldsmith 1990) are increasingly finding application in experimental systems (cf. Kornai 1991; Carson-Berndsen 1993; Kirchhoff 1996; Church 1987b,a) because of their richness and their more direct relation to the acoustic signal, in contrast to phonemic representations. However at the lexical level, they can generally be calculated relatively easily from the more compact, but less general, phonemic representations. Because of the widespread use of phonemes, the concept is discussed in more detail below; for fuller explanations, textbooks on phonology should be consulted.

The central question in phonological lexical representation, in cases where the notion of phoneme alone is not fully adequate, is that of the *level of representation* (*level of description*, *level of abstraction*). There are three main levels, each of which is an essential part of a full description, and which needs to be evaluated for all but the simplest applications, *morphophonemic*, *phonemic*, and *phonetic*, which are characterised below.

MORPHOPHONEMIC: The morphophonemic level provides a simplification of phonological information with respect to the phonological level; the simplifications utilise knowledge about the morphological structure of words, and permit the use of *morphophonemes*, (a near-synonym is *archiphoneme*) which stand for classes of morphologically and phonologically related phonemes.

A standard example of a morphophoneme is the final obstruent in languages with final obstruent devoicing, including Dutch and German. For example, the phonemic representation German Weg /ve:k/ 'way' – Wege/ve:gə/ 'ways' corresponds to a morphophonemic representation {ve:G} – {ve:G+ə}, which simplifies the description of the stem of the word. The morphophoneme {G} stands for the phoneme set {/k/, /g/}, and selection of the appropriate member of the set (the appropriate feature specification) is triggered by the morphological boundary and neighbouring phonological segments. Alternatively the morphophoneme may be said to consist of the underspecified feature bundle shared by /k/ and /g/, or more technically, the feature bundle which *subsumes* the feature bundles of /k/ and /g/.

An example from English is the alternation /f/ - /v/ in plural formation in one class of nouns, as in *knife* /naIf/ - *knives* /naIvz/, which can be represented morphophonemically as {naIV} - {naIV+z}. The morphophoneme {V} stands for the phoneme set {/f/, /v/}. Here, too, selection of the phoneme (specification of the underspecified subsuming feature bundle) is determined by the morphological boundary and the phonological properties of neighbouring segments.

A corresponding level is necessary for the description of spelling: cf. variations such as English *y*-*ie* in *city* - *cities*, or German *s*-*ss*- β as in *Bus* -

Busse, $Ku\beta$ (Kuss in the new orthography) – $K\ddot{u}sse$ and $Fu\beta$ – $F\ddot{u}\beta e$. Morphophonemic representations augmented by realisation rules are a useful *compression technique* for reducing lexicon size:

- Lexica can be stem-based, and thus have fewer entries, and all inflections can be automatically calculated by rule for any stem in the lexicon.
- Morphotactic and morphophonological rules can be used for extending lexica of fully inflected attested forms, and for checking such lexica for consistency.

For requirements such as these, the use of morphophonemic representations, supplemented by morphological construction rules and morphophonemic mapping rules is recommended (Koskenniemi (1983), Karttunen (1983), Ritchie et al. (1992), Bleiching et al. (1996) for descriptions of various practical approaches).

There are no standard conventions for the representation of morphophonemesmorphophonemics, whether computer readable or not (but see the SAMPA alphabet for French, Appendix B); capital letters are often used in linguistics publications. Note that this use of capital letters at the morphophonemic level should not be confused with the use of ASCII upper case codes in the SAMPA alphabet at the phonemic level.

Citations of *morphophonemic* representations are often delimited with brace brackets $\{\dots\}$.

PHONEMIC: The phonemic level is a standard intermediate level corresponding to criteria outlined in more detail below. The standard European computer readable phonetic alphabet is SAMPA (Appendix B): this alphabet is used for the main languages of the European Union, and is recommended for this purpose. The internationally recognised standard alphabet for phonemic representations is the International Phonetic Alphabet (IPA). The IPA alphabet is used for the most part in the text of this handbook, and is shown in Appendix A. One of the main functions of the International Phonetic Association since its inception over 100 years ago has been to coordinate and define standards for this alphabet.

Until relatively recently, the special font used for the IPA has made it difficult to interface it with spoken language systems, and for this reason a number of computer-readable encodings of subsets of the IPA have been made for various languages (cf. Allen 1988; Esling 1988, 1990; Jassem and Lobacz 1989; Ball 1991). The standard computer phonetic alphabet for the main languages of the European Union is the SAMPA alphabet, developed in the ESPRIT SAM and SAM-A projects (cf. Wells 1987, 1989, 1993b,a; Llisterri and Mariño 1993); see also Appendix B. SAMPA is widely used in European projects, both for corpus transcription and for lexical representations (see also the chapter on Spoken Language Corpora).

However, there is a standard numerical code for IPA symbols (cf. Esling (1988, 1990); Appendix B), and developments in user interfaces with graphical visualisation in recent years are leading to the increasing use of the IPA in its original form, particularly in the speech lab software which is used in

spoken language system development.

Citations of *phonemic* representations are standardly delimited by slashes $/\dots/$.

PHONETIC: At the phonetic level further details of pronunciation, beyond the phonemically minimal features, are given. Since the relation between the phonemic and the phonetic level can be described by general rules mapping phonemes to their detailed realisations (allophones) in specific contexts (cf. Woods and Zue 1976), it is strictly speaking redundant to include these regular variants in a lexicon. However, for reasons of efficiency, detailed phonetic word models for speech recogniser training or for speech synthesis may be calculated using phonological rules and stored. Essentially this is a software decision: whether to use tables (for efficiency of lookup) or rules (for compactness and generality) for a given purpose.

A specific version of the phonetic level of transcription is *phonotypic* transcription, defined as a mapping from the phonemic level using regular phonological rules of assimilation, deletion, epenthesis (cf. Autesserre et al. 1989); this level is frequently used for generating additional word models to improve speech recognition. Since the amount of phonetic detail which can be processed depends heavily on the vocabulary size and the number of phonological rules which are considered relevant, no general recommendation on this can be given.

There is no widely used standard ASCII encoding of the entire IPA for computer readable phonetic representations and therefore no recommendations can be given on this. A proposal by John Wells, the originator of SAMPA, is under discussion. Currently, individual laboratories use their own enhancements of phonemic representations. However, the fuller encodings mentioned in connection with the phonemic level of transcription are eminently suitable for interface purpose at the phonetic level, and will no doubt be increasingly used where more detailed phonetic information is required.

Citations of *phonetic* forms are standardly delimited by square brackets $[\ldots]$.

Chapters 4, 5, 3 should also be consulted in respect of levels and types of corpus representation.

6.4.3 Prosodic information

The area of word prosody, and, more generally, the description of other prosodic units which have quasi-morphemic functions, is gradually emerging as an important area for spoken language lexica. For present purposes, prosodic properties are defined as properties of word forms which are larger than phonemes. Further specification in phonetic terms (e.g. F0 patterning) and in semantic terms (e.g. attitudinal meaning) may also be given but is not essential for present purposes.

One type of lexical information on prosody pertains to phonological or morphological properties of words, such as Swedish pitch accents, or stress posi-

tions in words. Some aspects of word prosody are predictable on the basis of the regular phonological and morphological structure of words, but some are idiosyncratic. Examples in English where word stress is significant include the noun-verb alternation type as in *export* – /''ekspo:t/ (Noun), /eksp''o:t/ (Verb). In German, word stress is significant for instance in distinguishing between compound separable particle verbs and derived inseparable prefixed verbs as in *übersetzen* – /''y:bezetsən/ (compound) vs. /y:bez''ɛtsən/ (derivation).

It has been shown (cf. Waibel 1988) that taking word prosody into account in English can produce a significant improvement in recognition rate.

In addition, there is lexical information associated with prosodic units which occur independently of particular words, and therefore may themselves be regarded as lexical signs and be inventarised in a prosodic lexicon (cf. Aubergé 1992). To give a highly simplified example in a basic attributevalue notation, a prosodic lexicon for an intonation language might have the following structure.

This kind of information, in which prosodic categories function as a kind of *morpheme* with an identifiable meaning, is generally not regarded as lexical information, but treated as a separate layer of organisation in language. Intonation is being taken increasingly into account for *prosodic parsing* in two main senses of this term:

- 1. Analysis of speech signal in respect of the fundamental frequency (F0, F-zero) trajectory, for speech recognition, in relation to words, sentences and dialogue units.
- 2. Analysis of sentence structure for the generation of intonation patterns in speech synthesis.

Prosodic representation in the lexicon is in general restricted to the prosodic properties of words, such as stress position in English, Dutch, and German words, or tonal accent in Swedish words, or to rhythmically relevant units such as the syllable and the foot. For spoken language processing in which prosody plays a role, it is also necessary to include an inventory of prosodic forms, and their meanings, which play a role at the sentence level, independently of specific words: i.e. a prosodic lexicon.

It should be borne in mind that in linguistics, "prosody" currently has a broader meaning, and covers all properties of pronunciation which are not directly concerned with defining consonants and vowels. Prosody in this sense covers, for example, syllable structure and phonological word phonotactics, as well as the more traditional categories of intonation, accent, and phrasing.

The IPA defines symbols for representing lexical and non-lexical types of prosody, and a subset (for word prosody) has been encoded in the SAMPA alphabet. However, the state of knowledge in the area of prosody is less stable than in the area of segmental word structure, and a range of different conventions is available (cf. Bruce 1989); in this area, there are SAMPA "dialects", for instance replacing SAMPA " and % for primary and secondary stress by the more iconic ' (single quote) and '' (two single quotes) or " (double quote).

The ToBI (Tones and Break Indices) transcription, originally developed for American English, has now been applied to several languages (see also Chapter 5).

In oriented spoken language lexicography within the VERBMOBIL project, attribute-based formal representations of prosodic features in the lexicon have been developed using the ILEX (Integrated Lexicon) model and the lexical knowledge representation language DATR (cf. Bleiching 1992; Gibbon 1991).

There is an increasing tendency no longer to regard prosodic representations as totally exotic and quite unlike anything else. But there is still insufficient consensus on lexical prosodic features to permit generally valid recommendations to be made for prosodic representations in the lexicon. For most purposes, plain SAMPA or ToBI style symbols will be adequate. For covering new ground with extended lexica for use with discourse phenomena at the dialogue level, a lexical knowledge representation language with a more general notation, as illustrated above, may be more appropriate.

6.4.4 Recommendations on lexical surface information

- 1. Define the basic lexical entry type, as recommended in the previous section, and its notation.
- 2. Establish a machine-readable orthography convention.
- 3. Define the phonological level of representation (e.g. canonical phonemic).
- 4. Specify requirements in respect of word prosody representation.
- 5. Specify requirements in respect of phrasal and discourse prosody (intonation) information.
- 6. Use the European standard machine readable phonemic alphabet SAMPA (see Appendix B).
- 7. Select an appropriate system of prosodic representation, such as SAMPA, SAMPROSA (see Appendix B), INTSINT, or ToBI notation, where required.
- 8. Ensure that the relation between notations and representations used in the *lexical database* and the *system lexicon* are well-defined, and that they are

completely consistent with notations and representations in other resources such as *corpora* and in the different parts of the system, such as the *word lattice* produced by the speech recogniser, operated on by a *stochastic language model*, and further processed by a sentence parser.

6.5 Morphological information

6.5.1 Types of morphological information

Morphology is concerned with *generalisations* about words as lexical signs, in respect of surface form, meaning, distribution and composition. More generally, morphological information is information about semantically relevant word structure; the smallest morphological unit is the *morpheme*, often defined as the *smallest meaningful unit* in a language. Morphemes should not be confused with phonological units such as the *phoneme syllable* and its constituents, which are used for describing the structure of words from the point of view of their *pronunciation*, without direct reference to *meaning*. For applications of morphology to speech recognition see Althoff et al. (1996), Bleiching et al. (1996), Geutner (1995).

The domain of morphology may be divided in terms of the *functions* of morphological operations, i.e. inflectional *agreement* or *congruence* vs. *word formation*, or in terms of the *structures* defined by morphological operations, i.e. *affixation*, (*prefixation*, *suffixation*, *infixation* or *prosodic modification*) vs. compounding (*concatenation* of *stems* or *words*). These two dimensions can be represented as follows:

OPERATION	AGREEMENT	WORD FORMATION
AFFIXATION	Inflection	Derivation
STEM/WORD CONCATENATION		Composition

There is a gap in the table with regard to the use of stem or word concatenation for agreement; however, so-called *periphrastic constructions* with verbs, typically with auxiliary verbs and participles or infinitives, may be assigned to this slot, prepositions relate to nouns in a comparable way. Compare English John will come with French Jean viendra, or English Give it to the cook with German Geben Sie es dem Koch. English lacks an inflectional future, but has periphrastic (phrasal) modal or infinitive complement future forms such as John will come tomorrow, John is going to come tomorrow, as well as the present tense as a general or neutral tense form, as in John comes tomorrow (contrast with anecdotal narrative, such as "You know something? This morning Julie comes in and there's this pigeon sitting on her desk ..."). There are other intermediate cases which sometimes present difficulties in classification and where the solution is not always immediately obvious:

• Are the degrees of comparison of adjectives *inflections* or *derivations* (i.e. positive *loud*, comparative *louder*, superlative *loudest*)?

- Are participles *inflected forms* of verbs or *derivations* as deverbal adjectives?
- Are infinitives of verbs inflected forms or derivations (cf. They want an answer vs. They want to know; To be or not to be; He begged to come, He decided to come)?

Traditional treatments often treat these forms together with inflections, presumably because of their regularity and the involvement of suffixation. They are generally better treated as derivations, however, because they have different syntactic distributions from other inflections of the same stems, and may be additionally inflected as adjectives or nouns (cf. the orthographic form of the perfect participle in French with *être* verbs: *Elle est venue* – *She has come*.

6.5.2 Applications of morphology

Morphological structuring is useful for the following tasks:

- The treatment of large vocabularies for speech recognition and synthesis by means of rule-based generation of inflected forms from stems.
- The prediction of new (unattested, unknown) words for speech recognition on the basis of known principles of word composition, and known attested parts of words.
- Rule-based assignment of stress patterns.
- Word recognition by stem spotting.
- Construction of subword language models for speech recognition.

There are two main ways of structuring words internally into word subunits (word constituents):

- 1. SEMANTIC ORIENTATION. On morphological grounds, word forms may be decomposed into smaller meaningful units, the smallest of which are morphs, the phonological forms of morphemes; an intermediate unit between the morph and the word form is the stem.
- 2. PHONOLOGICAL ORIENTATION. On phonological grounds, word forms may be decomposed into smaller pronunciation units, the smallest of which are phonemes; an intermediate pronunciation unit is the syllable.

It is important to note that decomposition into syllables is not isomorphic with decomposition into morphs. For example, phonological has the syllable structure /fo . $n \partial \cdot l o \cdot d I$. $k \partial l$ and the morph structure /fon + $\partial + l \partial d I k$ + ∂l , which are quite different from each other.

In addition to phonological decomposition, in the written mode word forms may be decomposed into smaller spelling units, graphemes, each consisting of one or more characters. An intermediate orthographic unit is the orthographic break (orthographic syllable), which is in general only needed for splitting words at line-breaks and does not correspond closely to either syllable or morph boundaries but combines phonological, morphological and orthographic criteria.

It has already been noted that in many languages, syllables and morphs do not always coincide; morphs may be smaller than or larger than syllables. For the core requirements of speech recognition, in which a closed vocabulary of attested fully inflected word forms is generally used, morphological structuring is not necessary. Phonological structuring into syllables, demisyllables, diphone sequences or phonemes is widely used in order to increase statistical coverage and to capture details of pronunciation (cf. Browman 1980; Ruske and Schotola 1981; Ruske 1985).

A brief outline of the main concepts in morphology, as they affect spoken language lexica will be useful in developing spoken language lexica (for more detail a textbook in linguistics should be consulted, e.g. Akmajian (1984)):

MORPHOLOGY: Morphology is the definition of the composition of words as a function of the meaning, syntactic function, and phonological or orthographic form of their parts. The morphology of spoken language is fundamentally the same as the morphology of written language in respect of meaning, syntactic function, and the combinability of *morphemes*. It differs in respect of morphophonological alternations, which differ from spelling alternations, and word prosody (for instance word stress patterns). General definitions are given here; examples are given below.

Morphotactics (word syntax) is the definition of the composition of words as a function of the forms of their parts.

Inflection is that part of morphology which deals with the adaptation of words to their contexts within sentences: on the basis of agreement (congruence), e.g. between subject and verb.

Word formation is that part of morphology which deals with the construction of words from smaller meaningful parts.

Derivation is that part of word formation which deals with the construction of words by the concatenation of stems with affixes (prefixes and suffixes). Compounding (composition) is that part of word formation which deals with the construction of words by concatenating words or stems.

SIMPLE MORPHOLOGICAL UNITS: Traditional terminology varies in this area. A standard but incomplete definition of a *morpheme*, for instance, is that it is "the minimal meaning-bearing unit of a language". This definition is not entirely satisfactory, however, and for present purposes the sign-based model and the unit of *word* will be used as the starting point.

A morpheme is the smallest abstract sign-structured component of a word, and is assigned representations of its meaning, distribution and surface (orthographic and phonological) properties. More informally, morphemes are parts of words defined by criteria of form, distribution and meaning; i.e. they have meanings and are realised by orthographic or phonological forms (morphs). They have no internal morphological structure. Traditionally, the two main kinds of morpheme are;

Traditionally, the two main kinds of morpheme are:

• Lexical morphemes, characterised by membership of a large, potentially open class, with meanings such as properties and roles of objects, states and events. • *Grammatical morphemes*, characterised by membership of a closed class, defined by their distribution with respect to larger units such as sentences or complex words (e.g. inflectional and derivational endings; function words such as prepositions, articles).

Morphs are, in traditional linguistics, the orthographic or phonological forms (realisations) of morphemes. Orthographic morphs consist of graphemes (either single letters or fixed combinations of letters); in traditional phonology, phonological morphs consist of phoneme sequences with a prosodic pattern (e.g. word stress).

Roots or *bases* (lexical morphs) are the morphs which realise lexical morphemes and inflectable grammatical morphemes, and function as the smallest type of *stem* in *derivation* and *compounding*. *Affixes* (*prefixes*, *suffixes*) are morphs which realise the inflectional and derivational beginnings and endings of words.

A *free morph* is a morph which can occur on its own with no affixes or prosodic modifications as a separate word; a *bound morph* is a morph (generally an affix) which always occurs together with at least one other morph (typically a *stem* in the same word.

- COMPLEX MORPHOLOGICAL UNITS: The structure of words is, like the structure of sentences, defined recursively, since the vocabulary of a language (including new coinages) is potentially unlimited. The functional and formal classification of morphological word structure (compounding and derivation, see above) takes this into account. Where 'out of vocabulary words' are likely to be encountered, *morphotactic rules* and a *morphological parser* or *morphological generator* may be required in order to supplement the lexicon. The condition of recursive structure does not apply to inflection, which, given a finite set of stems, defines a finite set of fully inflected word forms (in agglutinative languages possibly an extremely large finite set):
 - INFLECTIONAL AFFIXATION: A word (fully inflected word) is a stem morphologically concatenated with a full set of inflectional affixes, e.g. English *algorithm* + s = algorithms or German ge + segn + et + en 'blessed' (plural participle or adjective).

DERIVATIONAL AFFIXATION: A stem is

- either a root (i.e. lexical morph), e.g. tree, algorithm
- or a stem morphologically concatenated with a derivational affix, e.g. algorithm + ic, algorithm + ic + al + ly, non + algorithm + ic + al + ly, etc.
- COMPOUNDING: A compound word is a word morphologically concatenated with a word or a stem.
- MORPHOPHONOLOGICAL AND ORTHOGRAPHIC ALTERNATIONS: The operation of *morphological concatenation* is defined for present purposes as "concatenation and modification of segments at morph boundaries by boundary phenomena." The details of pronunciation and spelling are altered in morphologically complex items. An example of morphophonological alternation is /f/ /v/ in *knife* /natf/ /natvz/. An example of orthographic alter-

nation is y - i - ie in fly, flier, flies. These alternants can be described by rules:

- 1. Morphophonological rules are rules (analogous to spelling rules) which describe morphophonological alternations, i.e. the differences between pronunciations of parts of composite words and pronunciations of corresponding parts of simplex words.
- 2. Spelling rules are rules which describe spelling alternations, i.e. the differences between spellings of parts of composite words and the spellings of corresponding parts of simplex words.

A standard technology for formulating spelling rules and morphophonological rules is *Two-Level Morphology* (cf. Koskenniemi (1983), Karttunen (1983); cf. Ritchie et al. (1992)).

6.5.3 Recommendations on morphology

- 1. Decide whether word subunits may have a role to play in the intended application.
- 2. For a large vocabulary (> 5000 words) spoken language lexical database of a highly inflecting language as a general resource, consider using a *morphology component* to generate fully inflected forms, either on demand or precompiled as a (very large) table.
- 3. For specific problem areas such as the identification of *new words* or increasing the robustness of a recogniser, consider using morphological units in speech recognition.
- 4. In highly inflecting languages, consider the use of stochastic language models based on word stems as an alternative to fully inflected words.

6.6 Grammatical information

6.6.1 Statistical language models

Language models are a major area of research and development, and crucial for the success of a speech recognition system. Chapter 7 should be consulted on this topic.

In speech recognition systems, the mapping of the digitised acoustic forms of words on to symbolic representations for use as lexical lookup keys is performed by stochastic speech recognisers, which may incorporate information about the phonological structure of a language to a greater or lesser extent, with *word models* for matching with the acoustic analysis of the speech signal. Details of standard practice can be easily be found in the literature (cf. Waibel and Lee 1990).

In written language processing, a comparable task is Optical Character Recognition (OCR), and in particular, handwriting recognition; there is no comparable task in conventional natural language processing or computational linguistics, where letters are uniquely identified by digital codes, and

dictionary access may be trivially optimised by encoding letter sequences into tries (letter trees, letter-based decision trees). However, in linguistic terms, in each case the task is the identification of word forms as lexical keys.

6.6.2 Sentence syntax information

Syntactic information is required not only for parsing into syntactic structures for further semantic processing in a speech understanding system, but also in order to control the assignment of prosodic information to sentences in *prosodic parsing* and *prosodic synthesis*.

Syntactic information is defined as information about the distribution of a word in syntactic structures. This is a very common, indeed "classical", but specialised use of the words "syntax" and "syntactic" to pertain to phrasal syntax, i.e. the structure of sentences. Other more general uses of the terms for linguistic units which are larger or smaller than sentences are increasingly encountered, such as "dialogue syntax", "word syntax" (for morphotactics within morphology).

Within this classical usage, the term *syntax* is sometimes opposed to the term *lexicon*; the term *grammar* is sometimes used to mean syntax, but sometimes includes both phrasal syntax and the lexicon.

Strictly speaking, a stochastic language model is a probabilistic sentence syntax, since it defines the distribution of words in syntactic structures. However, the notion of syntactic structure used is often rather elementary, consisting of a short fixed-length substring or window over word strings, with length two (bigram) or three (trigram). It is also used with quite a different function from the classical combination of sentence syntax and sentence parser.

Sentence syntax defines the structure of a (generally unlimited) set of sentences. Syntactic lexical information is traditionally divided into information about *paradigmatic* (classificatory; disjunctive; element-class, subclasssuperclass) and syntagmatic (compositional; conjunctive; part-whole) relations. The informal definitions of these terms in linguistics textbooks are often unclear, metaphorical and inconsistent. For instance, temporally parallel information about the constitution of phonemes in terms of distinctive features is sometimes regarded as paradigmatic (since features may be seen as intensional characterisations of a class of phonemes) and sometimes as syntagmatic (since the phonetic events corresponding to features occur together to constitute a phoneme as a larger whole). The relation here is analogous to the relation between intonation and sentences, which are also temporally parallel, and in fact treated in an identical fashion in contemporary computational phonology. From a formal point of view, this is purely a matter of perspective: the internal structure of a unit (syntagmatic relations between parts of the unit) may be seen as a property of the

unit (paradigmatic relation of similarity between the whole unit and other units). In lexical knowledge bases for spoken language systems it is crucial to keep questions of syntagmatic distribution and questions of paradigmatic similarity apart as two distinct and complementary aspects of structure.

The part of speech (POS, word class, or category) is the most elementary type of syntactic information. One traditional set of word classes consists of the following: Noun or Substantive, Pronoun, Verb, Adverb, Adjective, Article, Preposition, Conjunction, Interjection. POS classifications are used for tagging written corpora (texts or transcriptions), for the purpose of information retrieval or for the training of class-based statistical language models (Chapter 7); fairly standard POS tagsets have defined for a number of taggers (automatic tagging software; see the results of the EAGLES Working Group on machine readable corpora).

Two main groups of POS category are generally identified:

- 1. Lexical categories are the open classes which may be extended by word formation: Noun, Verb, Adjective, Adverb.
- 2. Grammatical categories are the closed classes which express syntactic and indexical relations: Pronoun and Article (anaphoric and deictic relations), Preposition (spatial, temporal, personal relations etc.), Conjunction (propositional relations), Interjection (dialogue relations).

The granularity of classification can be reduced by grouping classes together (this particular binary division is relevant for defining stress patterns for example) or increased by defining subcategories based on the *complements* (object, indirect object, prepositional or sentential object, etc.) of words (in various terminologies: their *valency* or *subcategorisation frames*, *case frames*, transitivity properties). For further information, introductory texts on syntax, e.g. Sells (1985) or Radford (1988) may be consulted.

In theoretical and computational linguistics, grammars are classified in terms of the *Chomsky hierarchy* of formal languages which they *generate* (i.e. define), and often represented as equivalent automata. Some aspects of this classification are discussed in connection with stochastic language models in Chapter 7. For further information, standard computer science compiler construction literature can be consulted.

6.6.3 Recommendations on grammatical information

- 1. Consult the chapter on language models for further details on the application of language models in speech recognition.
- 2. Consult the chapter on speech synthesis for further details on grammar in the production of speech.
- 3. Consult the results of the EAGLES Working Groups on Formalisms and on Computational Lexica for additional information about the use of grammatical information in sentence parsing.
- 4. Ensure that a thorough grammatical analysis is performed on the

application-relevant data in order to ensure that no more grammatical complexity is introduced than is actually needed.

6.7 Lexical content information

6.7.1 Lexical semantic information

The treatment of semantics in lexica for spoken language systems is less uniform than that of word forms, and tends to be highly application—specific:

- In single word control systems, a direct mapping to an operation or action is sufficient.
- In a dictation system, a mapping to text is needed, with or without deeper analysis.
- Conversely, in a text-to-speech system, a full semantic analysis is also not generally necessary.
- In a database front end application, a translation into a database query language is required.
- In a spoken language understanding or translation system sophisticated kinds of semantic representation such as artificial intelligence representation languages (semantic networks or frame structures), or more general formalisms such as first order predicate calculus or typed attribute logics are required.

An outline of some of the basic lexical semantic relations between lexical units was given at the beginning of this section in the context of the discussion of relations between lexical units (lexical relations).

Reference should be made to the results of the EAGLES Computational Lexica Working Group for further information on lexical semantics, and to standard texts such as Lyons (1977) or Cruse (1986).

6.7.2 Pragmatic information

The central area of application for pragmatic information from the point of view of spoken language is in dialogue situations, in which prosody (intonation, emphasis, accentuation) is required in addition to word-based information: prosodic information is typically associated with speaker-centred pragmatic information such as topic focussing, speaker attitudes, and dialogue turn-taking. For further details of dialogue structure, Chapter 13 and the results of the EAGLES Working Group on Corpora should be consulted. Because spoken language corpora are generally highly application oriented, and therefore bound to a particular speaker or set of speakers, with relatively homogeneous properties of register, speech style, and dialect, including pronunciation, vocabulary, grammar, and intentions with respect to specific actions, the kind of pragmatic information required for particular lexical items is also restricted.

Spoken language lexica differ in this way from large-scale general coverage lexica, though the need for such lexica in the spoken language area is growing. In many respects spoken language lexica differ considerably from text-based written language lexica in respect of pragmatic information. The kinds of pragmatic information required are generally limited to information about a few speech act types (question, answer, instruction, etc.). The advent of complex spoken language dialogue systems makes treatment of unorthodox items such as discourse particles, including hesitation markers, and of word fragments a necessity.

6.7.3 Idiomatic information

Pragmatic idioms such as greetings are often treated as holistic lexical items, i.e. as "canned text", and included in the lexicon in full. The same applies to fixed idioms such as *Come to think of it*, ... in the meaning 'I just thought of another relevant point, namely ... '; formulations like *I will just come to think of it*, ..., *Come to consider it*, ... etc. do not have idiomatic meaning.

The most complex problems arise in the case of idioms with variant forms, such as *If you twist my arm, then* ..., i.e. 'If you give me a really good reason, then ...', where as a joke forms such as *Twist my arm!* or even *would you mind twisting Fred's arm?* have the idiomatic meaning. However, formulations such as *Don't twist my arm!*, *It was my arm that Tony twisted* etc. tend to have the literal, rather than the idiomatic meaning.

Idioms may be associated with a specific range of prosodic patterns (intonation patterns; for example, in *How do you do?*, the words *how*, the first occurrence of *do* and *you* might be emphasised as a joke, but not with the standard pragmatic idiom connotation. In the example given above, if the word *twist* is given a noticeably stronger accent than *arm*, the same also applies.

From a more general perspective, so-called *functional units* (sequences of functional words which behave as a phonological unit) and *clitics* (functional words which combine with lexical words to form a sequence which behaves as a phonological unit), share a number of properties with idioms in the more traditional sense of the term. Characteristic of these units is that they have special phonological characteristics, with deletion and assimilation of segments to their neighbouring environment.

An example of a functional unit in English is $I c'n \ldots /aIkn/for I can /aIkm/ in informal, fast speech or particularly unstressed contexts.$

An example of a clitic in English is he's /hi:z/ or even /hIz/, for he is /hi: Iz/.

Some cliticised sequences have become *lexicalised* (i.e. independent lexical items) in informal styles, e.g. *can't* /ka:nt/ and *cannot* /k{not/ for *can not* /kæn no/.

A general solution to the lexical treatment of idioms is not currently available either for written language processing or spoken language processing, and further research and development is needed in this area in view of the frequency of idioms in actual corpora.

6.7.4 Recommendations on semantic information

- 1. Ensure that the application domain is as precisely specified and modelled as possible.
- 2. Determine the size of vocabulary which can be processed and the types of semantic information required by the domain model, and possibly consider ways in which to minimise the vocabulary needed for this domain.
- 3. Consult the results of the EAGLES Working Group on Computational Lexica (this is an area in which Spoken Language and Written Language processing overlap considerably).

6.8 Lexicon structure

The area of *lexicon structure* deals with the organisation of information in lexica. Models for lexical information, and types of lexical information, are dealt with in the preceding sections. Terminology varies considerably in this area. The structure of a spoken language lexicon may be seen from the following points of view:

- LEXICAL FORMALISMS, LEXICON REPRESENTATION LANGUAGES: Representation conventions of various types (symbolic notations, programming languages, database languages, logical formalisms, purpose-designed knowledge representation languages), which are suitable for formulating *lexical models*.
- LEXICON ARCHITECTURE: The choice of basic objects and properties in the lexicon, and the structure of the lexicon as a whole, such as a table of items, a trie (decision tree), an inheritance hierarchy, a semantic network, a database.

6.8.1 Spoken language lexicon formalisms

Spoken language lexicon formalisms (representation languages) may be broadly classified according to their use:

- 1. Linguistically and phonetically based working notations.
- 2. Implementation languages for the operational phase.
- 3. Algebraic and logical formalisms for formal definition.

Where an ad hoc solution is required for a very small lexicon, and where lexicon structure is simple, a lexicon may be written directly in a standard programming language suitable for high-speed runtime applications, traditionally Fortran but more recently C, or in a higher level language such as LISP or Prolog. Recent developments are moving towards high level

knowledge representation languages which are specifically designed to meet all three of the above criteria equally well, in that they are useful working notations, have efficient implementations, and are formally well-defined. Some of these are also used for general written language lexica. A more

detailed classification of formal representation systems may be given as follows:

- 1. General data structures (lists, tables or matrices, tree structures designed for optimal lexical access).
- 2. Programming languages (C for efficiency; LISP or Prolog for flexibility).
- 3. Database systems.
- 4. General text markup languages such as SGML.
- 5. Knowledge representation languages (inheritance networks, semantic networks, frame systems).
- 6. Linguistic knowledge representation languages, commonly based on attribute-value logics.
- 7. Lexical knowledge representation languages (attribute based inheritance formalisms) such as DATR.

General data structure definitions for these representations are required for developers and for theoretical work on the complexity and efficiency of lexica and lexicon processing. Standard textbooks on data structures and algorithms should be consulted for this purpose.

Conventional programming languages are generally used for performance reasons in runtime systems. They may also be used to implement small or simple lexica directly, in particular for rapid prototyping of these; this is not optimal software development practice, however, and not to be recommended for developing large or complex lexica, in particular those with highly structured linguistic information.

Database management systems (DBMSs) are widely used for general lexical resource management, including large-scale lexica with rich information which needs to be accessed flexibly and efficiently (see Appendix H). In the SAM project, an ORACLE database management concept for spoken language corpora and lexica was developed (cf. Dolmazon et al. 1990).

General text markup languages are used for integration with large, preanalysed written corpora in the development of natural language processing systems and in statistical basic research in computational linguistics, but so far have not been used in spoken language system development (cf. the results of the EAGLES Working Group on Text Corpora). Implementations of SGML are readily available.

Knowledge representation languages (KRLs) are used for developing complex semantic and conceptual knowledge representations, and for integrating spoken language front ends with knowledge based systems; see Schröder et al. (1987), Sagerer (1990); more generally, cf. Bobrow and Winograd (1977), Brachman and Levesque (1985), Charniak and McDermott (1985), De Mori et al. (1984), Young et al. (1989).

Linguistic formalisms in general are discussed in the results of the EAGLES Working Group on Grammar Formalisms, which should be referred to in this connection.

Lexical knowledge representation languages (LKRLs) are a relatively new development. They are coming to be used in knowledge acquisition for integrated lexica which contain a variety of complex lexical information from phonology through morphology and syntax to semantics and pragmatics. They provide a means of bridging the gap between complexity of lexical information and easy-to-read representations using sign-based lexicon models. A LKRL which has been used in several natural language processing and language and speech projects is DATR (cf. Evans and Gazdar 1989; Cahill 1993; Cahill and Evans 1990; Andry et al. 1992; Gibbon 1991, 1993; Bleiching 1992; Langer and Gibbon 1992). This is the language which is used for basic attribute-value representations in this chapter. A number of public domain DATR implementations are available and can be obtained from the Web sites.

6.8.2 Lexicon architecture and lexical database structure

Lexicon architecture pertains to the *choice of basic objects and properties* in the lexicon, and to the *overall structure* of the lexicon. More formally, it defines the relation which assigns lexical properties to lexical entries.

The term "architecture" generally refers to the structure of system lexica, but the term is also justified in connection with lexical database structure, particularly when more complex relational or object-oriented structures are concerned.

The basic objects in terms of which an architecture may be defined were discussed in the section on lexical information for spoken language.

The overall structure of a spoken language lexicon is determined by a range of declarative, procedural and operational criteria such as the following:

- The complexity of the information assigned to lexical entries.
- The complexity of the relations defined between lexical entries.
- The particular subset of objects and properties defined for a given application lexicon.
- Linguistic and logical compression techniques such as redundancy rules or, more generally, inheritance hierarchies.
- Task driven directionality of access.
- Variety of information required for access (from phonological to pragmatic).
- Performance requirements of software (including lingware) size and speed of access.
- Techniques of acquisition and maintenance (with respect, for instance, consistency).
At the one extreme is the ideal notion of a fully integrated sign-based model with non-redundant specification of entries and property inheritance; in between is the efficient database management system used for large scale lexical databases (see Appendix H on DBMSs), and at the other extreme is the simple pronunciation table which is the starting point for the training of speech recognition devices.

The choice of lexicon architecture on the basis of parameters such as those listed above, and taking into account practical constraints from the actual working environment, is application specific. There is no single principle of organisation which applies to all lexica.

The closest approximation to a neutral form of spoken language lexicon organisation is a sign-based general background lexicon organised as a database with flexible access. Such a lexicon is basically knowledge acquisition oriented, and can function as a source for the specialised lexica required for different speech synthesis and recognition applications. Specialised models for sublexica which are optimised for particular applications can then be formulated, and sublexica can be automatically compiled out of the main lexicon into application-specific notations and structures.

The organisation of a lexicon determines the general properties of the formalism to be used with the lexicon. Conversely, available formalisms determine tractable forms of lexicon organisation in terms of data structures, algorithms and programming environments (cf. Knuth 1973; Carbonell and Pierrel 1986; Rudnicky et al. 1987; Lacouture and Normandin 1993). Object-oriented system architectures, with local encapsulation of all aspects of representation and processing, permit the construction of hybrid systems with functionally optimised components; by analogy, the lexicon itself can be conceived as a hybrid object system if required.

This is in effect the situation in current speech recognition technology, in which a more or less large set of HMMs representing words, for instance, can be seen as a procedurally sophisticated lexicon with acoustically driven lookup of keys which are then used to access the main lexicon. Although the standard perspective is to see the two components as separate, they can be seen as objects which are both located in hybrid spoken language system spoken language system lexicon components.

Current research on new object-oriented interactive incremental spoken language system architectures raises many new questions about the role of a lexicon. One major question is whether the lexicon is an object (or system of objects) in its own right, or whether the lexicon is distributed over the system components and is thus a *virtual lexicon*, or which components of the system, e.g. morphology and word semantics, or sentence parsing and propositional semantics, interact directly. Questions such as these are the subject of ongoing basic research, and it would be premature to make specific recommendations at this point. For a broader discussion of lexicon architectures, the work of the EAGLES Working Group on Computational Lexica should be consulted.

6.8.3 Lexicon architecture and the structure of lexical databases

The architecture of a lexicon, in particular of a lexical database, is determined partly by the types of declarative knowledge it contains, partly by considerations of access and interaction with other databases or modules. The main features of spoken language lexical databases have already been discussed. In practice, a spoken language lexical database is often a set of loosely related simpler databases (e.g. pronunciation table, signal annotation file, stochastic word model, and a main lexical database with syntactic and semantic information), and is part of a larger complex of databases involving speech signal files, transcription files (orthographic and phonemic), and annotation (labelling) files which define a function from the transcriptions into the digitised speech signal. However, in the interests of consistency it is helpful to take a more general lexicographic point of view, and to see a lexical database for spoken language development as a single database, in which relations between lexical items and their properties at all levels, from acoustics through word structure to syntax, semantics and pragmatics are defined.

The major problem in deciding how to organise a lexical database is the *ambiguity* of word forms. In a spoken language system, the focus is on the *pronunciation*, i.e. on *phonemic word forms* (not the orthography, though this is often used as a conveniently familiar form of representation). The key issue here is *homophony*, i.e. a phonemic word form associated with at least two different sets of lexical information, and thus logically involving a *disjunction* in the database.

In a simple traditional database model based on fixed-length records, in which each field represents a specific attribute of the entity which the record stands for, there is a record for each lexical entry associated with a homophone, uniquely identified by a serial number. However, for specific applications such as the training of a speech recogniser it is convenient to have just one record for each word form. In a database which is optimised for this application, the disjunction required by the homphone is *within a single record*, rather than distributed over alternative records which share the same value for the pronunciation attribute. Structures of this type are typically used in pronunciation dictionaries (pronunciation lexica, pronunciation tables) for speech recognition. Disjunctive information of this kind within the lexical database corresponds to non-deterministic situations and the use of complex search algorithms in actual spoken language systems.

6.8.3.1 A simple database type: Pronunciation tables

Pronunciation tables (pronunciation dictionaries) hardly correspond to the intuitive concept of a lexical database, which implies a fairly high degree of complexity, but they are nevertheless a useful source of practical examples of a simple lexical database structure.

Pronunciation tables define the relation between orthographic and phonemic representations of words. Often they are defined as functions which assign pronunciations (frequently a set of variant pronunciations) to orthographic representations; this is an obvious necessity for text-to-speech lexica, but in speech recognition applications in which orthographic transcriptions (which are easier to make and check than phonemic transcriptions) are mapped to phonemic representations for the purpose of training speech recognisers, the use of a pronunciation table of this type is relevant.

Boundaries				
morpheme: +				
word:	#			
liaisonless group:	##			
phonological syntagma: § (in phrasal entries)				
Phonemes (in IPA or SAMPA notation), including				
a notation for the French archiphonemes.				
Phonological diacritics				
latency mark	"			
(for consonants pronounced in liaison contexts				
or morphological linking)				
consonant deletion mark ' (e.g. for final consonants)				

Table 6.3: Frequently used symbols

A pronunciation table which involves pronunciation variants (see below) provides a simple illustration of the orthographic noise problems, represented by disjunctions in the database.

Pronunciation tables have to fulfil a number of criteria, in particular the criterion of unambiguous notation, of consistency with orthographic transcriptions and other transcriptions of a particular corpus, and of simple and fast processing.

General proposals for the interchange of lexical information about word forms, including morphological, phonological and prosodic information, have been made for different languages. They do not have standard status at the current time, but they are sufficiently similar to justify recommendation. A standard for French has been described (cf. Pérennou and De Calmès 1987; Autesserre et al. 1989), containing the features tabled in

Boundaries					
morpheme:	+				
stem-inflection boundary:	#+				
word in compounds:	#				
word in phrases:	##				
syllable:					
primary stress:	,				
secondary stress:	'' (two single quotes)				
Additional co	nventions				
Additional co The boundaries # and ## a	nventions re both				
Additional co The boundaries # and ## a coextensive + and . boun	nventions re both daries.				
Additional cc The boundaries # and ## a coextensive + and . boun Where + and . boundaries	nventions re both daries. are coextensive,				
Additional cc The boundaries # and ## a coextensive + and . boun Where + and . boundaries . is written before +.	onventions re both daries. are coextensive,				
Additional cc The boundaries # and ## a coextensive + and . boun Where + and . boundaries . is written before +. The stress marks ' and ''	nventions re both daries. are coextensive,				
Additional cc The boundaries # and ## a coextensive + and . boun Where + and . boundaries . is written before +. The stress marks ' and '' are written immediately b	onventions re both daries. are coextensive, before the vowel, not				

Table 6.4: VERBMOBIL diacritics

Figure 6.3.

For the spoken language lexicon in the German VERBMOBIL project the same basic principle has been adopted (cf. Bleiching and Gibbon 1994), with extensions for incorporating prosodic information, as in Table 6.4. Table 6.5 shows an extract from the VERBMOBIL pronunciation table in the

VERBMOBIL WIF (Word form Interchange Format) convention; following current practice, it is organised according to orthographic keys.

The convention has been designed to permit the removal of information which is not required, or the selection of useful subsets of the table using simple UNIX tool commands; the use of "1" for primary and secondary stress permits simple generalisation over both.

6.8.3.2 More complex lexical databases

In a complex project, lexical information from several sources may need to be integrated in a fashion which permits flexible further development work even when the information cannot easily be reduced to a logically fully consistent and well-defined system. A situation such as this will arise when alternative modules, based on different principles, are to be made available for the same system. For instance, two different syntactic components will define different forms of syntactic ambiguity and be associated in different ways with semantic ambiguities. And morphological ambiguities arise with inflected forms in highly inflecting languages. In order to achieve any kind of

ASCII orthography	Extended SAMPA transcription		
Angst	?'aNst		
Annahme	?'an#n''a:.m+@		
Apparat	?a.pa.r'a:t		
April	?a.pr'Il		
Aprilwoche	?a.pr'Il#v"O.x+@		
Arzttermin	?'a6tst#tE6.m''i:n		
Aschermittwoch	?''a.S6#m'It#v''0x		
Auf_Wiederh"oren	?aUf##v'i:.d6#h''2:.r+@n		
Auf_Wiederschauen	?aUf##v'i:.d6#S''aU.+@n		
Auf_Wiedersehen	?aUf##v'i:.d6#z''e:.+@n		
August	?aU.g'Ust ?'aU.gUst		
Augustwoche	?aU.g'Ust#v''O.x+@		
Ausweichm"oglichkeit	?'aUs#v"aIC#m"2:k.+lIC.+kaIt		

Table 6.5: Extract from the VERBMOBIL pronunciation table

integration, at least the word form representations will need to be consistent. The hybrid information sources will have to be represented as conjunctions of the values of independent attributes (i.e. fields within a record), with separate disjunctions, where relevant, within fields.

In general, spoken language projects have been based on the idealised notion of a *single*, *well-defined*, *consistent* and *complete*; this situation might reasonably be expected to correspond to the reality of a system developed in a single laboratory at one specific location. However, larger scale projects need to be able to cope with hybrid lexical information of the kind just outlined. A project of this type is the VERBMOBIL project funded by the German government, with international participation.

A general product–oriented solution would obviouly use a product standard database (see the Appendices and Chapter 5), but an illustration of the typical R&D style UNIX database is given here for the sake of simplicity as an example of a database structure designed for hybrid lexical information.

- 1. Internal database structure (standard UNIX database format):
 - database: header records followed by body records
 - header: header_record_1 header_record_2 header_record_3
 - body: body_record_1 ... body_record_n
 - header_record_1: (record containing attribute names, i.e. field names)
 - header_record_2: (record defining internal conjunctive/disjunctive structure of attribute values, i.e. field contents)
 - header_record_3: (record containing source of information)

- body_record_i: (record containing values for a given entry)
- 2. Example of record structure:
 - Header: (the designations A3 etc. refer to projects delivering particular types of information)
 RECORD 1: Orth A3 B1 C1 D1

```
RECORD 2: Orth A3.Phon B1.Wortart,B1.Kasus,B1.Genus,B1.Num,
B1.Detagr,B1.Definit,1.Semobj,B1.Semattr
C1.Syncat1_C1.Syncat2 D1.Syncat
RECORD 3: reference.ort a3joha.lex b1naeve.lex c1jung.lex
d1peters.lex
```

- Body: Mutter mU!t6 nomen_akk,fem,sg,@empty@,@empty@,Raute,@empty@; nomen,nom,fem,sg,@empty@,@empty@,Raute,@empty@ Nom,OBJEKTTYP nom
- Note that the spaces designate conjunction (i.e. field separators), while the semicolons designate disjunction
- 3. Example of human-readable formatting

```
Entry 372: Mutter

Orth: Mutter

A3: mU!t6

B1: nomen,akk,fem,sg,@empty@,@empty@,Raute,@empty@

nomen,nom,fem,sg,@empty@,@empty@,Raute,@empty@

C1: Nom,OBJEKTTYP

D1: nom
```

On UNIX systems, laboratory-specific acquisition and access routines for ASCII lexical databases are frequently writen with sandard UNIX script languages for ASCII character stream processing. If the resources are available to produce fully specified C and C++ programmes, then of course this is to be preferred. The UNIX tools are useful for prototyping and *ad hoc* format conversion and informal exchange within the speech development community, but are not to be recommended for commercial use.

The following example illustrates simple UNIX script programming for human-readable format conversion (transformation of selected named attributes of a database record into the attribute format given above):

```
#!/bin/sh
# dbviewr
# Prettyprint of single entries
# and attributes in lexicon database
# with regular expression matching
# Uses UNIX tools:
# gawk (i.e. gnu awk), sed, tr
```

```
# (Note: sed and tr are used for illustration, and would
        normally be emulated in gawk)
#
# Database structure:
# Header: Record 1: Fields containing attribute names.
         Record 2: Other information.
#
# Body: Records >2: Database relation.
if [ $# -1t 3 ]
 then
 echo "Usage: dbview dbname attribute* regexp"
 exit
fi
# The GNU version of the awk script language is used:
gawk '
# Transfer the keyword from the command line to an awk variable:
BEGIN {keyword = ARGV[ARGC-1]}
# Identify the attributes in the first record whose values
# are to be queried.
NR == 1 {{for (i=2 ; i < ARGC ; i++)
        {for (j=1 ; j <= NF ; j++)
        if (ARGV[i] == $j) {attrib[j] = "yes"; attname[j] = $j}}}
        {for (i = 2 ; i < ARGC ; i++)
        ARGV[i]=""}}
# Find required keyword entry/entries in body of database,
# print required values and set 'found' flag:
$1 ~ keyword && NR > 2 {print "\nEntry " NR-2 ":", $1
        {for (i=1 ; i <= NF ; i++)
        if (attrib[i] ~ "yes") {print " " attname[i] ":\t" $i
                                found="yes"}}}
{last=NR}
# Print message if no entry was found for the keyword.
END {if (found!="yes") {print "No entry found for",keyword,
"in", ARGV[1]}}
' $* |
# Pipe to sed script language,
# translate all sequences of two colons into a slash, all single
# colons into a single colon followed by eight spaces:
sed -e "s/;;/\//g
        s/;/&
                     /g" |
# Pipe to tr character translator,
# translate all single colons into a linefeed (newline):
tr ";" "\012"
```

For an overview of related format conversion techniques, see Aho et al.

(1987), Dougherty (1990), Wall and Schwartz (1991).

6.8.4 Recommendations on lexicon structure

- 1. If sufficient resources are available, and the scale of the development warrants it, a commercial database system should be preferred.
- 2. The development of new re-usable resources in an R&D context is likely to be best supported by the use of a LKRL with an efficient and flexible implementation, and the use of formalisms in this category can therefore be recommended for development purposes where re-usable resources are concerned.
- 3. Whatever the choice of formalism for a spoken language lexicon, the interface between the lexicon and other components of a spoken language system needs to be given close attention.
- 4. A given single formalism is highly unlikely to provide optimal features for the development and construction of all components of a given system in the near future.
- 5. Object-oriented database structures and system architectures are likely to be used increasingly in spoken language systems in order to permit the encapsulation of module-specific optimal formalisms, implementations and information.
- 6. Weigh the advantages of the use of specialised commercial databases against the practical convenience and portability of ASCII database standards associated with the UNIX operating system and UNIX ASCII processing tools for database format conversion and database access.
- 7. Consider whether a system lexicon should be a single component, or, as is more usual in speech recognition applications, a "distributed lexicon" with lexical information associated with the *speech recognition*, *stochastic language model* and *lexical search* components.

6.9 Lexical knowledge acquisition for spoken language

The most general declarative perspective on a spoken language lexicon which is required for lexicon acquisition is that of the "omniscient lexicographer": the lexicographer "knows", in principle, all the possible categories and relations which may hold between a sign, its meaning and pronunciation, and other signs; all properties of a sign have equal status in terms of possible access. This idealised view, while useful for general lexical databases, is not appropriate for the construction of more specialised spoken language lexica for the classical types of spoken language system, unless these are derived from a more general lexical database as sublexica.

6.9.1 Stages in lexical knowledge acquisition

A prerequisite for lexical acquisition is to define the following items:

- An application dependent lexicon model which defines relevant types of lexical information.
- A lexical database model which defines the storage structure for lexical information.
- A characterisation of the sources of lexical data (e.g. orthographic transcriptions, other databases; manual information provided systematically or *ad hoc* by a lexicographer)
- Definition of procedures for constructing a lexicon from the data on the basis of these models and sources.
- Definition of procedures for validating the consistency of the database.
- Definition of procedures for extracting complex lexical information (e.g. coverage statistics for attributes and attribute combinations).
- Tools for implementing the procedures.

The logistics involved in these procedures are entirely non-trivial in a large project or consortium, and especially so with a complex spoken language understanding system or speech-to-speech translation system.

6.9.2 Types of knowledge source

The types of lexical knowledge source for a spoken language system depend largely on the application. There are few general sources of lexical spoken language material (for instance with pronunciation and general frequency information) for any language. The construction of such a source is a major task which requires concerted action on a large scale by specialists of a whole language engineering community. It is a formidable task for many theoretical and practical reasons, but nevertheless one which will require a great deal of effort in the coming years. The two major sources of lexical knowledge for spoken language lexical systems are:

- 1. existing dictionaries (to some extent),
- 2. application specific corpora (to a large extent),
- 3. results of descriptive, theoretical and computational linguistics (to some extent).

There is still a definite lack of general resources in the area (cf. the introduction to this chapter), and the construction of application-derived, generalisable resources will be a major task for any project and for the entire spoken language community in the coming years.

General lexical material is required for the lexical knowledge in general coverage text-to-speech systems, as well as for broad application pronunciation tables for speech recognition.

6.9.2.1 Dictionaries

Useful sources of information are generally available dictionaries, particularly pronouncing dictionaries, provided that they adhere to accepted standards of consistency and expressiveness of notation, and are available in electronic form. An overview of some sources was given at the beginning of this chapter, and reference should be made to the results of the EAGLES Working Group on Computational Lexica for further examples.

6.9.2.2 Corpora

Spoken language lexica are application specific, and necessarily so when corpus-derived frequency information is needed. An example of a corpusderived lexicon type for speech recognition was given above. Another type of corpus-derived lexicon is the diphone word list widely used in speech synthesis technology; for this, phoneme label alignment with the speech signal is required, with the aid of which diphones are defined in the signal for further processing. The chapter on Spoken Language Corpora contains detailed information on procedures of corpus treatment, and the results of the EAGLES Working Group on Text Corpora should also be consulted.

6.9.2.3 Acquisition tools

At the current state of the art, there are few generally available tools for constructing spoken language lexica, either by extraction from existing dictionaries or from corpora. Lexicon construction usually takes place "in house" in individual laboratories or project consortia; lexicon formats consequently vary greatly.

For information on general acquisition tools in the sense of lexicographers' work benches, reference should be made to the results of the EAGLES Working Group on Computational Lexica. It is not appropriate in this context to go into the vast domain of Machine Learning and its application to the (semi-)automatic acquisition of lexica from data.

Of greatest practical use for the development of spoken language lexica in the area of word forms are the tools required for creating different kinds of word form list and word form table from corpora; the general parameters associated with acquiring syntactic, semantic and pragmatic information are not unique to spoken language lexica (though the details, for instance of spoken language dialogue, indeed differ greatly from spoken to written language).

It is a common practice is either to write custom-made programmes in C, or, where speed of processing is not at a premium, to use standard UNIX script languages for processing orthographic transcriptions. Neither of these procedures is particularly difficult, because of the relatively straightforward and well-understood procedures and associated algorithms.

The simplest approach for many applications where processing time is not critical, for instance with small lexica, or where batch-style processing is acceptable, is to use UNIX tools such as *grep*, *tr*, *sed*, *uniq*, *cut*, *tail*, *spell* and *awk*. For descriptions of these tools, a UNIX manual or textbook, or the man page on-line information on a UNIX system should be consulted;

techniques for specific database oriented UNIX tools are described by Aho et al. (1987), Dougherty (1990), Wall and Schwartz (1991).

An example of database formatting was given above. Simple examples of UNIX tool applications are illustrated in grossly simplified form below in order to convey an idea of the sort of corpus pre-processing required for ASCII-based spoken language lexicon acquisition.

• Orthographic transcription to word list:

```
#!/bin/sh
# Simple wordlist generator
echo Wordlist generator
tr -sc 'A-Za-z' '\012' < $1 | sort | uniq > wordlist.srt
echo Wordlist in file 'wordlist.srt'
```

• Orthographic transcription to frequency list:

```
#!/bin/sh
# Simple word frequency generator
echo Word frequency generator
tr -sc 'A-Za-z' '\012' < $1 | sort | uniq -c > wordlist.frq
echo Wordlist in file
'wordlist.frq'.
```

• Orthographic transcriptiontranscription!orthographic to digram frequency table:

```
#!/bin/sh
# Simple digram table generator
echo Digram generator
tr -sc 'A-Za-z' '\012' < $1 > lines.txt
tail +2 lines.txt > tailed.txt
paste lines.txt tailed.txt | sort | uniq -c > digrams.tab
echo Digram frequency table in file 'digrams.tab'.
```

Digram frequency information of this type is the basis for the construction of statistical *language models*.; this simple illustration is, however, not to be compared with state of the art technology (cf. Chapter 7).

6.9.3 Recommendations on lexicon construction

- 1. Identify and acquire the data resources needed for construction of the lexical database and, in the second instance, the system lexicon.
- 2. Construct a basic wordlist from the available data and decide on intensional and extensional lexical coverage for the lexical database and the system lexicon.
- 3. Design the lexicon database to be acquired so as to maximise the information available for the development of the system lexicon.

- 4. Consider establishing a UNIX tool library for convenient ASCII format conversion in database construction, transformation and transfer, in addition to using standard database software.
- 5. Results of the EAGLES Working Group on Computational Lexica should be consulted for the design of lexicon models and lexical resources for written language; some of the techniques are also applicable to spoken language lexica and lexical databases.
- 6. Do not underestimate the time and effort needed for constructing and validating software for lexicon construction and lexical lingware: lexicon construction is labour intensive.
- 7. If possible, use a commercial database management system (DBMS), and consult Appendix H.

In this area, there is an urgent need for the development of standardised lexicon construction tools and the provision of lexical resources.

6.10 Outlook

In this chapter, an overview of applications, requirements, lexical information, lexical representation and lexicon structure for spoken language systems is given, with particular reference to the differences between lexica designed for spoken language as opposed to written language processing. Specific recommendations were provided at the relevant points.

The state of the art for spoken language lexica is such that at the level of word forms, many basic standard techniques have been developed and are widely used. Large-scale spoken language lexical resources are required for current research and development with both statistical and knowledgebased technologies, and need to function as. reference sources of standard, stylistic and regional pronunciations, and in general vocabularies which are characteristic of spoken language, including, for instance, discourse particles,

Spoken language lexical resources in the form of actual lexical databases and tools for constructing them are, however, sadly lacking – even more so than for written language, partly because of the specific complexities of spoken language, and partly also to the fact that the construction of spoken language lexical databases is highly labour-intensive, therefore expensive, and consequently proprietary and in-lab developments are not often made generally available.

The major requirement in this area is simply that this gap urgently needs to be filled. Local and national project work can go some way to meeting the need. In view of the labour-intensive character of lexicon construction and, the feasibility of constructing lexical representation languages and lexical acquisition tools which can be deployed in multifunctional and multilingual environments, cooperative international initiatives are essential.

An important area which has not been touched on is the question of how

spoken language lexica relate directly to the human user. This question may be answered in terms of two main areas:

1. Cognition and the spoken language lexicon.

2. Ergonomics and the spoken language lexicon.

The first answer is a theoretical perspective, pertaining to the theoretical sense of the mental lexicon, lexical access and lexicon acquisition strategies of the user, which the lexicon of a spoken language system, in some sense, models (cf. Marslen-Wilson 1989).

The second answer touches on a more practical area, namely how the role of the lexicon in a spoken language system can be defined in such a way as to permit the most ergonomic, user-friendly development and deployment of complex spoken language systems. The former concept is an essential source of ideas and criteria for the latter.

Treatment of these areas, taking extensive work in experimental psychology and psycholinguistics as well as user requirements into account, would be premature at present, but may be predicted to play a prominent role in future research and development.

7 Language models

7.1 Introduction

7.1.1 Goals of the chapter

This chapter gives an overview of language modelling in the context of large vocabulary speech recognition and covers the following topics:

- Why do we need language modelling in a speech recognition system, and what are the particular problems?
- Stochastic language models are introduced to capture the inherent redundancy of the language subset relevant for the specific recognition task.
- The definition of so-called perplexity or, to be more exact, corpus perplexity is introduced as a quantitative measure for the language constraints that depend on both the chosen language model and the test corpus.
- The need for smoothing techniques in language modelling is discussed. Smoothing will be based on the so-called leaving-one-out technique which can be viewed as a special type of cross-validation.
- To illustrate the specific problems in language modelling, we consider smoothing methods for the widely and successfully used bigram and trigram language models.
- For smoothing, three techniques are presented in detail: linear discounting, linear interpolation and absolute discounting.
- A detailed description of a full trigram model is given. As an extension, the combination of a trigram model with the so-called cache model is considered. Experimental results are discussed for the Wall Street Journal (WSJ) corpus along with practical issues in the implementation of language models.
- We describe refinements over the standard trigram model by using word classes that are automatically learned. We also discuss grammar based language models. In particular, so-called link grammars are able to capture long range dependencies as opposed to the short range dependencies modelled by trigram models.
- In a recognition system, the language model is used during the decision process, which is usually referred to as search in speech recognition. To illustrate the integration of the language model into the recognition procedure, we study two search techniques in detail, namely the search for the single best sentence and the generation of word graphs or lattices.
- The advantage of the word graph is that the acoustic recognition can be decoupled from postprocessing steps like applying the language model or the dialogue module. Here significant progress has been achieved so that now reliable and yet compact word graphs (lattices) can be generated.
- In a *final note* (Section 7.8), we describe the mathematical details of the smoothing techniques, in particular the EM algorithm.

The primary application we consider is large vocabulary speech recognition with applications like text dictation and automatic dialogue systems. Some of the techniques presented are maybe useful in other applications, too, like systems for voice commands and guided dialogues, where a finite state network might be sufficient as language model. For most non-experts and maybe even the experts in speech recognition, it still is a surprise that the trigram language model performs as well as it does. In contrast, grammar based language models (i.e. models based on linguistic grammars) are far from being competitive at the present time. Therefore, the description focusses on the trigram model and related issues such as the sparse data problem and smoothing. This chapter is only able to touch upon some of the issues in language modelling. For other overviews, see Jelinek (1991); Jelinek et al. (1991a, 1992). For related topics such as the use of stochastic methods for language acquisition and language understanding, see Gorin et al. (1991) and Pieraccini et al. (1993), respectively.

7.1.2 Advice to the reader

The material presented in this chapter is relevant both to corpus characterisation and to design and evaluation issues in automatic speech recognition. The chapter is different from most of the other chapters:

- Language models as such are an important part of the whole recognition system, and therefore this chapter will give more details of the techniques and algorithms.
- To fully understand the problems and implications of language modelling, it is necessary to consider some of the statistical and mathematical problems involved.
- The sections are written in such a way that, at first reading, the reader may skip the mathematical parts.
- Most of the mathematical details and derivations have been put in an appendix (see 7.8).

The reader should consult also the following chapters:

- spoken language lexica,
- dialogue,
- assessment methodologies and experimental design,
- assessment of recognition systems.

7.2 System architecture for speech recognition

7.2.1 Bayes decision rule

Every approach to automatic speech recognition is faced with the problem of taking decisions in the presence of ambiguity and context, and of modelling the interdependence of these decisions at various levels. If it were possible to recognise phonemes (or words) with a very high reliability, it would not be necessary to rely heavily on delayed decision techniques, error correcting techniques and statistical methods. In the near future, this problem of reliable and virtually error free phoneme or word recognition without using high-level knowledge is unlikely to be solved for large-vocabulary continuous-speech recognition. As a consequence, the recognition system has to deal with a large number of hypotheses about phonemes, words and sentences, and ideally has to take into account the "high-level constraints" as given by syntax, semantics and pragmatics. Given this state of affairs, statistical decision theory tells us how to minimise the probability of recognition errors (Bahl et al. 1983).

The word sequence $w_1...w_N$ to be recognised from the sequence of acoustic observations $x_1...x_T$ is determined as that word sequence $w_1...w_N$ for which the posterior probability $Pr(w_1...w_N | x_1...x_T)$ attains its maximum. The sequence of acoustic vectors $x_1...x_t...x_T$ over time t = 1...T is derived from the speech signal in the preprocessing step of acoustic analysis. Statistical decision theory leads to the so-called Bayes decision rule, which can be written in the form:

$$\underset{w_1...w_N}{\arg\max} \quad \{Pr(w_1...w_N) \cdot Pr(x_1...x_T | w_1...w_N)\}$$
(7.1)

where $Pr(x_1...x_T|w_1...w_N)$ is the conditional probability, given the word sequence $w_1...w_N$, of observing the sequence of acoustic vectors $x_1...x_T$ and where $Pr(w_1...w_N)$ is the prior probability of producing the word sequence $w_1...w_N$. The application of the Bayes decision rule to the speech recognition problem is illustrated in Figure 7.1.

The decision rule requires two types of probability distribution, which we refer to as stochastic knowledge sources, along with a search strategy:

- The language model, i.e. $Pr(w_1...w_N)$, is independent of the acoustic observations; its task is to incorporate restrictions on the way in which the words of the vocabulary can be concatenated to form whole sentences.
- The acoustic-phonetic model, i.e. $Pr(x_1...x_T|w_1...w_N)$, is the conditional probability of observing the acoustic vectors $x_1...x_T$ when the speaker utters the words $w_1...w_N$. Like the language model probabilities, these probabilities are estimated during the training phase of the recognition system. For a large vocabulary system, there is typically a set of basic recognition units that are smaller than whole words. Examples of these so-called subword units are phonemes, demisyllables or syllables. Often, context dependent phoneme units are also used, for example so-called triphones, i.e. phoneme units in a triphone context. The word models are then obtained by concatenating the subword models according to the phonetic transcription of the words in a pronunciation dictionary. In most systems, the acoustic-phonetic models are based on Hidden Markov models (Levinson et al. 1983; Bahl et al. 1983).



Figure 7.1: Bayes decision rule for speech recognition

• The decision on which spoken word have most probably been recognised is taken by maximising the product of the probabilities of the language model and of the acoustic-phonetic model over all word sequences. In such a way, the search strategy combines information and constraints coming from the different knowledge sources: the language model and the acousticphonetic model which comprises the set of basic subword units and the pronunciation dictionary. The optimisation procedure typically requires a search through a state space that is defined by the knowledge sources.

7.2.2 Stochastic language modelling

The task of a language model is to express the restrictions imposed on the way in which words can be combined to form sentences. In other words, the idea is to capture the inherent redundancy that is present in the language, or to be more exact, in the language subset handled by the system. This redundancy results from the syntactic, semantic and pragmatic constraints of the language and may be modelled by probabilistic or non-probabilistic ("yes/no") methods. For a vocabulary of three words A, B, C, Figure 7.2 illustrates the situation. What is always possible is to arrange the word

sequences in the form of a tree. Figure 7.2 shows the sentence tree for all four-word sentences. Now some of these word sequences may be impossible, some may be possible, and others may be very typical according to syntactic and semantic and maybe pragmatic constraints. The task of the language model is now to express these constraints by assigning a probability to each of the sentences. In simple cases like voice command applications, it might be sufficient to just remove the illegal sentences from the diagram and compress it into a finite state network.

For large vocabulary recognition tasks, such methods cannot be used because we have to allow any word sequence types, which is difficult to describe deterministically. The task of a stochastic language model is to provide estimates of these prior probabilities $Pr(w_1...w_N)$. Using the definition of conditional probabilities, we obtain the decomposition:

$$Pr(w_1...w_N) = \prod_{n=1}^{N} Pr(w_n | w_1...w_{n-1})$$
(7.2)

Strictly speaking, this equation requires a suitable interpretation of the variable N, the number of words. When considering a single sentence, the number of words is a random variable itself, and we need an additional distribution over the sentence lengths. In practice, the problem is circumvented by applying the above equation to a whole set of sentences and extending the vocabulary by a special symbol (or "word") that marks the end of a sentence (and the beginning of the next sentence).

For large vocabulary speech recognition, these conditional probabilities are typically used in the following way (Bahl et al. 1983). The dependence of the conditional probability of observing a word w_n at a position n is modelled as being restricted to its immediate (m-1) predecessor words $w_{n-m+1}...w_{n-1}$. The resulting model is that of a Markov chain and is referred to as *m*-gram model. The following types of model are quite common:

- trigram model: $Pr(w_n|w_1...w_{n-1}) = p(w_n|w_{n-2}, w_{n-1})$
- bigram model: $Pr(w_n|w_1...w_{n-1}) = p(w_n|w_{n-1})$
- unigram model: $Pr(w_n|w_1...w_{n-1}) = p(w_n)$ zerogram model: $Pr(w_n|w_1...w_{n-1}) = p(w_n)$

Here we have used W to denote the vocabulary size. Note that the zerogram model is a special unigram model with uniform probabilities. Other types of language models will be considered later in this chapter. The probabilities of these models are estimated from a text corpus during a training phase. However, due to the experimental conditions, we are faced with a particular problem that is usually referred to as the problem of sparse training data. We consider this problem in more detail. For bigram and trigram models, most of the possible events, i.e. word pairs and word triples, are never seen



Figure 7.2: Illustration of the decision problem for a three-word vocabulary

in training because there are so many of them. We have to make sure that nevertheless these unseen events are assigned a probability greater than zero. Otherwise word sequences that contain these unseen word bigrams or trigrams cannot possibly be hypothesised or recognised during the speech recognition process.

It is obvious that, apart from speech recognition, language models are also essential for optical character recognition (Mori et al. 1992) and language translation (Berger et al. 1994). It is interesting to mention that similar m-gram techniques are used in the context of acoustic-phonetic modelling. The main difference is the level at which the statistical data are collected, i.e. at the level of phonemes or the level of phones, which are the acoustic realisations of the phonemes. Phone bigrams and trigrams are referred to as diphones and triphones, respectively. Statistical techniques related to those used in language modelling can also be applied to language understanding (Pieraccini et al. 1993).

7.2.3 Types of language models

To illustrate the broad range of language model types, we mention some typical examples:

- no or uniform language model: Here, the idea is to use the same probability for all events; events can be either the words of the vocabulary or the sentences, if the number of sentences is limited. If all words are equiprobable, there is an implied model for the duration of a sentence: a sentence of N words then has a probability $(1/W)^N$.
- finite state language model: The set of legal word sequences is represented as a finite state network (or regular grammar) whose edges stand for the spoken words, i.e. each path through the network results in a legal word sequence. To make this approach correct from a probabilistic point of view, the edges have to be assigned probabilities.
- *m-gram language models*: In *m*-gram language models, all word sequences are possible, and the probability of the word predicted depends only on the (*m*-1) immediate predecessor words (see above).
- grammar based language models: Typically, these models are based on variants of stochastic context free grammars or other phrase structure grammars.
- other types: There are language models that make use of still other concepts like CART (classification and regression trees) (Breiman et al. 1984; Bahl et al. 1989) and maximum entropy (Lau et al. 1993; Rosenfeld 1994).

It should be noted that this classification of language models is not exhaustive, and a specific language model may belong to several types.

7.2.4 Why does the trigram language model work?

In order to show how the trigram works in recognition, F. Jelinek (Jelinek 1991) has given the example shown in Table 7.1.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	The	are	to	know	the	issues	necessar y	role	and	the	n ext	be	meeting
2	This	will		have	this	problems	data	thing	from			two	months
3	One	the		understand	these	th e	information	that	in				years
4	Two	would		do	problems		above	to	to				meetings
5	А	also		get	any		other	contact	are				to
6	Three	do		the	a		time	parts	with				week
7	Please	need		use	problem		people	point	where				days
8	In			provide	them		operators	for	requiring				
9	We			insert	all		tools	issues	still				
61				4.4.4			4.4.4		being				
62				1.1.1			4.4.4		during				
63				1.1.1			4.4.4		I				
64				1.1.1			1.1.1		involved				
65				1.1.1			1.1.1		would				
66				1.1.1			1.1.1		within				
				1.1.1			111						
93				request			factors						
94				respond			facts						
95				supply			1						
96				write			jobs						
97				me			MVS						
98				resolve			old						
636							mailroom						
637							marketplace						
638							provision						
639							reception						
640							shop						
641							important						

Table 7.1: Effect of trigram model on recognition (Jelinek 1991)

The spoken sentence was: "We need to resolve all the important issues within the next two months." The figure shows all the words that are assigned a probability higher than the word actually spoken for a trigram language model. The probabilities are based on using the words actually spoken as conditioning events. For example, for the two predecessor words "all the", the most likely word is "necessary", whereas the actually spoken word "important" is only in position 641. From the figure, it can be seen that function words like prepositions and articles tend to be better predicted than the content words. The reason is that the function words occur more often in a corpus and thus their trigrams are more reliable. At the same time, the function words are more difficult to recognise from the acousticphonetic point of view due to the coarticulation effects. So we see that there is an interesting symbiosis of trigram language models and acoustic-phonetic models. When the acoustic-phonetic model tends to be poor, as for function words, the trigram model tends to be strong. When the trigram model is weak as for content words, the acoustic-phonetic models are more reliable because content words are long and are less subjected to coarticulation.

7.3 Perplexity definition

7.3.1 Formal definition

Strictly speaking, to evaluate the quality of a stochastic language model, we would have to run a whole recognition experiment. However, as a first approximation, we can separate the two types of probability distribution in Bayes' decision rule and confine ourselves to the probability that the language model produces for a sequence of (test or training) words w_n , $n = 1, \ldots, N$. To normalise this prior probability with respect to the number N of words, we take the Nth root and take the inverse to obtain the so-called corpus (or test set) perplexity (Bahl et al. 1983):

$$PP := [Pr(w_1...w_N)]^{-1/N}$$
(7.3)

Inserting the decomposition into conditional probabilities of Eq.(7.2) and taking the logarithm, we obtain:

$$\log PP = -\frac{1}{N} \sum_{n=1}^{N} \log Pr(w_n | w_1 \dots w_{n-1})$$
(7.4)

To avoid confusions, we prefer the term "corpus perplexity" because it can be used for both training and test data. The above equations show that the corpus perplexity is the geometric average of the reciprocal probability over all N words. Apart from the constant factor (-1/N), the corpus perplexity is identical to the average conditional probability or likelihood. Therefore minimising the corpus perplexity is the same as maximising the log-likelihood function.

The perplexity measures the constraints expressed by the language model. From the viewpoint of the recognition task, we can say that the language model reduces the number of word choices during the recognition process. Thus the perplexity can be interpreted as the average number of word choices during the recognition process. As a first approximation, the perplexity measures the difficulty of a recognition task: the smaller the perplexity, the lower the error rate. For example, depending on the application and the language model, a recognition system with a vocabulary of 1000 words can have such strong language constraints that the recognition task is easier than digit recognition. This was true for all of the early speech recognition systems like HARPY and HEARSAY (Lea 1980). A special aspect in the definition of corpus perplexity should be noted. If a word in the corpus is assigned a probability of zero by the language model, the perplexity will be infinitely large. This is one of the real challenges for the language model: the prediction of the next word should be as good as possible without excluding any of the words of the vocabulary.

In some articles, the authors relate perplexity to entropy as used in information theory (Bahl et al. 1983). There the assumption is that the underlying

probability distribution of the language model is exactly known. However, for practical comparisons, so-called test set perplexity or corpus perplexity is more useful.

7.3.2 Implications of the perplexity definition

We consider in more detail the implications of the formal definition of the perplexity:

- Perplexity refers to written (e.g. transcribed) forms of the language only and completely ignores the acoustic-phonetic modelling. This may be viewed as a strength and a weakness at the same time.
- Perplexity is based on the written form of the spoken words or, to be precise, the fully inflected word forms; in speech recognition, there is a convention to call every sequence of characters between blanks a *word*.
- Perplexity requires a closed vocabulary. If a word occurs that is not part of the vocabulary, the perplexity definition may run into problems because it becomes infinitely large. This out-of-vocabulary word problem will be considered below.
- Perplexity is merely a single averaged scalar-valued quantity; there is no information about local variations across the corpus. It would be straightforward to define the variance; an even more informative method would use a histogram over local probabilities, i.e. reciprocal perplexities.
- By definition, perplexity depends on both a specific corpus and a specific language model. So it has a dual function: perplexity is a measure for characterising both the corpus *and* the specific language model. In other words, using the same language model, we can compare the difficulty of two corpora, i.e. their redundancy from the viewpoint of the language model. This also works the other way round: Using the same corpus, we can compare the quality of two language models.

The definition of perplexity involves the issue of *coverage* at several levels and in different aspects:

- *vocabulary coverage*: The vocabulary is assumed to be closed, i.e. each word spoken in the test set must be part of the vocabulary of the recogniser specified beforehand. In recognition tasks like text dictation, this problem is often circumvented by adding the *out-of-vocabulary* word to the conventional vocabulary.
- *bigram and trigram coverage*: The language model should cover those word bigram and word trigrams that are typical of the test sentences.
- *coverage measure*: The perplexity can be used as a quantitative measure of the coverage of the language model, i.e. the perplexity measures how well the language model covers the test sentences.

In most cases, the definition of the recognition vocabulary is based on the collection of representative text corpora. The most frequent words in the corpus define the recognition vocabulary. This method seems to be widely

used for recognition systems working in speaker independent mode. For speaker dependent systems, it is not practical to collect a sufficiently large corpus from a single person. Therefore, typically, some combination with a speaker independent corpus is used. There are special techniques that have been developed for this purpose of vocabulary personalisation (Jelinek et al. 1991a).

7.3.3 Extensions of the perplexity definition

What we have considered so far could be called word-based perplexity because the smallest units are the words of the language. This word based perplexity can be extended in two directions. First, we can go from written words to letters (characters) and define a letter based perplexity. The advantage then is that the set of symbols is guaranteed to be closed because it is simply the alphabet, i.e. the set of characters, and at the same time, the vocabulary is unlimited because any word can be made up from the symbols of the alphabet. The additional complication, however, is that the constraints of the spelling dictionary must be taken into account in addition to the usual language model constraints.

The second direction could be to go one step further and consider the set of phonemes as set of basic units. In this case, the pronunciation constraints (or in its more general form the phonotactic constraints of the language) would be taken into account in the perplexity definition. Such a phoneme based perplexity could measure all the constraints that are considered to be "prior" to the observed acoustic signal. For example, it is a well known fact that the difficulty of a recognition task depends very much on the phonetic similarities of the words to be recognised. In particular, the lengths of the words to be recognised play an important role. As a first approximation, this could be taken into account by normalising word based perplexity with respect to the number of phonemes per spoken word. However, even in today's research systems, we are still limited to the word based perplexity.

7.3.4 Recommendations

When comparing the difficulty of two recognition tasks, the perplexity is only a first approximation. For such a purpose, it is helpful to remember that there are a number of important details to be checked. This is even more true when a direct comparison of two language models for the *same corpus* is performed. In comparing perplexities, the following points should be checked:

- 1. What is the exact vocabulary, and above all, what is the exact size of the vocabulary?
- 2. How are punctuation marks and in particular sentence boundaries treated? Often, in text dictation, punctuation marks are included in the vocabulary.

- 248 Part II: Spoken language characterisation
- 3. How is the unknown or out-of-vocabulary word handled? Is it included in the calculation of the perplexity, or is the perplexity calculated only by averaging over the *spoken* words?
- 4. What are the conventions for representing numbers and dates?
- 5. It makes a difference whether the probabilities or their logarithms are averaged. To avoid potential confusions, the corpus perplexity should be computed for the corpus as a whole. If it is computed on a sentence-bysentence basis, it should be done by using the log-perplexities rather than the perplexities themselves.

7.4 Language model smoothing: modelling unseen events

7.4.1 Problem formulation

To illustrate the fundamental problem in language modelling, we consider a specific example, namely a bigram model. Conventional methods like the widely used maximum likelihood method (Lehmann 1983) produce the relative frequencies as estimates for the bigram probabilities:

$$p(w|v) = \frac{N(v,w)}{N(v)}$$

Here, vw is the word bigram under consideration, and N(v, w) and N(v) are the numbers of observed word bigrams (vw) and words v, respectively. Now assuming a vocabulary size of W = 20000 words, there are $W^2 = 400$ million possible word bigrams, but the training corpus consists rarely of more than 10 million words. As a result, the conventional probability estimate for each unseen event is zero, and only 2.5 % of all bigrams can be observed due to the size of the training set. For trigram language models, the effect will be even more disastrous.

To overcome these shortcomings of the conventional probability estimates, we have to apply some sort of "smoothing" to the relative frequencies and to make sure that each probability estimate is larger than zero. It is important to understand that the need for language model smoothing results from very specific boundary conditions:

- All word combinations should be possible, i.e. there is no word sequence with exactly zero probability.
- For such models, the amount of training data is always small, even if the size is 100 million of running words.
- To take into account unseen events, i.e. events that were not seen in the training data, we make use of cross-validation (Efron and Tibshirani 1993).

In this section, we describe suitable methods for smoothing, namely linear discounting with backing-off, linear interpolation and absolute discounting

with backing-off. The parameters of these models will be estimated by leaving-one-out (Duda and Hart 1973), which is a special type of cross-validation.

Throughout the chapter, we will use the so-called counts to describe the training data $w_1...w_n...w_N$. For example, trigram counts N(u, v, w) are obtained by counting how often the particular word trigram (uvw) occurs in the training data:

$$N(u, v, w) = \sum_{n: \ (w_{n-2}w_{n-1}w_n) = (uvw)} 1$$

Other counts are defined in the same way. We use the following notation for counts, where the dot symbol can be interpreted as a sum or wild card:

N(h, w): number of joint events (h, w); $N(\cdot, w) = N(w)$: number of joint events (\cdot, w) for a fixed word w; $N(h, \cdot) = N(h)$: number of joint events (h, \cdot) for a fixed history h; $N(\cdot, \cdot) = N$: total number of joint events (h, w).

In addition, we have to count how often a certain count r has occurred. These "count-counts" are defined as follows:

 $N_r(h, \cdot)$: number of events (words) w that occurred exactly r times for a fixed history h;

 $N_r(\cdot, \cdot)$: total number of joint events (h, w) that occurred exactly r times.

Events with counts with r = 0, 1, 2 occur often and have special names:

- r = 0: unseen events, i.e. events never observed in the training data;
- r = 1: singleton events, i.e. events that were observed exactly once;
- $r=2{\rm :}$ doublet on events, i.e. events that were observed exactly twice.

In the following, we formulate the smoothing techniques by considering two types of history. As an example, we consider a trigram language model:

- a specific history h; e.g. h = (u, v) for a trigram (uvw);
- a generalised history \overline{h} ; e.g. for trigram models with (h, w) = (u, v, w), typically we define (\overline{h}, w) to be the bigram $(\overline{h}, w) = (v, w)$.

7.4.2 Linear discounting and backing-off

We start with linear discounting in the so-called backing-off model (Katz 1987; Jelinek 1991). This model is exceptional in that it allows virtually closed-form solutions in conjunction with leaving-one-out. Thus we can dis-

cuss all the important formulae of the approach without recourse to numeric calculations. The model is defined by the equation:

$$p(w|h) = \begin{cases} (1 - \lambda_h) \cdot \frac{N(h, w)}{N(h, \cdot)} & \text{if } N(h, w) > 0\\ \lambda_h \cdot \left[\frac{\beta(w|\overline{h})}{\sum\limits_{w':N(h, w') = 0} \beta(w'|\overline{h})} \right] & \text{if } N(h, w) = 0 \end{cases}$$
(7.5)

It is easy to verify that the probabilities p(w|h) sum up to unity. In particular, λ_h is the total probability mass to be distributed over the unseen words for a given history h:

$$\lambda_h = \sum_{w:N(h,w)=0} p(w|h)$$
(7.6)

Here we have two types of parameter sets to be estimated:

- the discounting parameters λ_h for each history h;
- the backing-off distribution $\beta(w|\overline{h})$ for a generalised history \overline{h} . Note that for each history h the generalised history \overline{h} must be well defined in order to have a backing-off distribution $\beta(w|\overline{h})$.

The unknown parameters in the discounting models will be determined with the so-called leaving-one-out method. This method can be obtained as an extension of the cross-validation or held-out method (Duda and Hart 1973; Efron and Tibshirani 1993), where the training text is split into two parts: a "retained" part for estimating the relative frequencies and a "held-out" part for estimating the optimal interpolation parameters. Here, the training text is divided into (N-1) samples as the "retained" part and only 1 sample as the "held-out" part; this process is repeated N times so that all Npartitions with 1 "held-out" sample are considered. The basic advantage of this approach is that all samples are used both in the "retained" part and in the "held-out" part and thus very efficient exploitation of the training text is achieved. In particular, singleton events are reduced from 1 to 0 so that the effect of not seeing events is simulated.

The application of leaving-one-out to the problem under consideration is as follows. By doing some elementary manipulations as shown in the appendix, we can decompose the log-likelihood function into two parts, one of which depends only on λ_h and the other depends only on $\beta(w|\overline{h})$:

$$F(\{\lambda_h\},\{\beta(w|\overline{h})\}) = F(\{\lambda_h\}) + F(\{\beta(w|\overline{h})\})$$

+ const($\{\lambda_h\},\{\beta(w|\overline{h})\})$ (7.7)

The λ_h dependent part is:

$$F(\{\lambda_h\}) = \sum_{h} \left[N_1(h, \cdot) \log \lambda_h + [N(h, \cdot) - N_1(h, \cdot)] \log (1 - \lambda_h) \right]$$
(7.8)

Taking the partial derivatives with respect to λ_h and equating them to zero, we obtain the closed-form solution:

$$\lambda_h = \frac{N_1(h, \cdot)}{N(h, \cdot)} \tag{7.9}$$

The same value is obtained when we compute the probability mass of unseen words in the training data for a given history h:

$$\sum_{w:N(h,w)=0} p(w|h) = \frac{N_1(h,\cdot)}{N(h,\cdot)}$$
(7.10)

The same value is obtained when using the Turing-Good formula (Good 1953; Nadas 1985; Katz 1987; Ney and Essen 1993). This formula is extremely useful when we want to estimate the total probability for events that were not seen in the training data. This formula tells us simply to take the number of singleton events, i.e. $N_1(h, \cdot)$, divide it by the total number of events, i.e. $N(h, \cdot)$, and use this fraction as estimate for the probability of *new* events. This rule can be applied at various levels such as single words (unigrams), word bigrams and word trigrams. For example, to check the coverage of the vocabulary, we can use this formula to estimate the probability with which we will encounter words that are not part of the vocabulary, i.e. that we have not seen in the training data.

For the backing-off distribution, the standard method is to use the relative frequencies as estimates:

$$\beta(w|\overline{h}) = \frac{N(\overline{h}, w)}{\sum\limits_{w'} N(\overline{h}, w')}$$
(7.11)

However, as shown in Generet et al. (1995), the leaving-one-out method can be applied to this task, too. In the appendix, we show how an approximate solution is obtained for this case. The resulting estimation formula is:

$$\beta(w|\overline{h}) \cong \frac{N_1(\overline{h}, w)}{\sum\limits_{w'} N_1(\overline{h}, w')}$$
(7.12)

where $N_1(\overline{h}, w)$ is defined as:

$$N_1(\overline{h}, w) = \sum_{h \in \overline{h}: N(h, w) = 1} 1$$
(7.13)

For convenience, we have chosen the normalisation $\sum_{w} \beta(w|\overline{h}) = 1$. This type of backing-off distribution will be referred to as singleton backing-off distribution or singleton distribution for short. Considering bigram modelling as a specific example, we see that the singleton backing-off distribution takes into account such word bigrams as "has been", "United States", "bona fide" etc., where the second word is strongly correlated with the immediate predecessor word. In these cases, the conventional unigram distribution tends to overestimate the corresponding probability in the backing-off distribution.

7.4.3 Linear interpolation

In linear interpolation, a weighted average between the relative frequencies $N(h, w)/N(h, \cdot)$ and the general distribution $\beta(w|\overline{h})$ is computed:

$$p(w|h) = (1 - \lambda_h) \frac{N(h, w)}{N(h, \cdot)} + \lambda_h \beta(w|\overline{h}) \quad \text{with} \quad \sum_w \beta(w|\overline{h}) = 1$$
(7.14)

In other words, the difference to linear discounting in connection with backing-off is that the more general distribution $\beta(w|\overline{h})$ is used in all cases (Jelinek and Mercer 1980; Nadas 1984). The mathematical framework becomes rather complex for estimating the unknown parameters in linear interpolation. In most cases, the so-called EM algorithm is used as described in the appendix; EM stands for expectation-maximisation.

7.4.4 Absolute discounting and backing-off

The basic idea is to subtract a constant from all counts r > 0 and thus, in particular, to leave the high counts virtually intact. The intuitive justification is that a particular event that has been seen exactly r times in the training data is likely to occur r - 1, r or r + 1 times in a new set of data. Therefore, we assume a model where the counts r are modified by an additive offset. From the normalisation constraint, it immediately follows that this must be a negative constant since the unseen events require a non-zero probability. Experimental results in Ney and Essen (1993) show that the resulting estimates are close to estimates obtained from the Turing-Good formula after suitable smoothing (Good 1953; Nadas 1985). We define the model for absolute discounting:

$$p(w|h) = \left\{ \begin{array}{cc} \frac{N(h,w) - b_h}{N(h,\cdot)} & \text{if } N(h,w) > 0\\ b_h \cdot \frac{W - N_0(h,\cdot)}{N(h,\cdot)} \cdot \frac{\beta(w|\overline{h})}{\sum\limits_{w':N(h,w')=0} \beta(w'|\overline{h})} & \text{if } N(h,w) = 0 \end{array} \right\}$$
(7.15)

We do the same manipulations as for linear discounting, i.e. separating the singletons, ordering and carrying out the sums. For $\beta(w|\overline{h})$, we obtain the same equation as for linear discounting. For the b_h dependent part, we obtain the following leaving-one-out log-likelihood function:

$$F(\{b_h\}) = \sum_{h} \left[N_1(h, \cdot) \log b_h + \sum_{r>1} r N_r(h, \cdot) \log [r - 1 - b_h] \right]$$
(7.16)

Taking the partial derivatives with respect to b_h , we obtain the following equation after separating the term with r = 2:

$$\frac{N_1(h,\cdot)}{b_h} - \frac{2N_2(h,\cdot)}{1-b_h} = \sum_{r\geq 3} \frac{rN_r(h,\cdot)}{r-1-b_h}$$
(7.17)

For this equation, there is no closed-form solution. However, there are upper and lower bounds. As shown in the appendix, we have the upper bound:

$$b_h \le \frac{N_1(h, \cdot)}{N_1(h, \cdot) + 2N_2(h, \cdot)}$$
(7.18)

and the lower bound:

$$\frac{N_1(h,\cdot) - \frac{1}{4}[N(h,\cdot) - N_1(h,\cdot) - 2N_2(h,\cdot)]}{N_1(h,\cdot) + 2N_2(h,\cdot)} \leq b_h$$
(7.19)

7.4.5 Conclusion

In this section, the leaving-one-out technique and the different smoothing techniques have been applied to the estimation of probabilities in stochastic language modelling only. However, it should be stressed that the techniques presented are by no means limited to language modelling. They may be useful in similar areas, where probabilistic models are applied and the number of free parameters to be trained is large. Examples of such fields are probabilistic grammars, machine learning, machine translation, information retrieval and database queries, probabilistic reasoning in expert systems, and any other field in which information is processed, decisions are taken and the sparseness of training data must be explicitly taken into account. For noiseless source coding, see the textbook by Bell, Cleary and Witten (Bell et al. 1990, p. 144).

7.5 Multilevel smoothing for trigram models

7.5.1 Problem formulation

When smoothing a trigram model with a bigram model, we have to keep in mind that the backing-off distribution itself requires smoothing. So the

bigram itself is smoothed by a unigram which again may be smoothed by a zerogram. Thus, we can define the following levels for a trigram event (u, v, w):

- the trigram level $\overline{h}_3 = (u, v)$, which defines the relative trigram frequencies as the level to start with;

- the bigram level \$\bar{h}_2 = (\cdot, v)\$;
 the unigram level \$\bar{h}_1 = (\cdot, \cdot)\$;
 the zerogram level \$\bar{h}_0\$ if the unigram estimates are unreliable.

It is helpful to explicitly write down the notation used in the following, in particular the definitions of the so-called singletons and the unseen events:

N(u, v, w): number of observations for trigram uvw; $N(u, v, \cdot) = N(u, v)$: number of observations for bigram uv; $N_0(u, v, \cdot)$: number of unseen trigrams starting with uv; $N_1(\cdot, v, w)$: number of trigram singletons ending in vw; $N_1(\cdot, v, \cdot)$: number of trigram singletons having v in the middle.

The definitions at the bigram and unigram level are similar:

 $N(u, \cdot) = N(u)$: number of observations for unigram u; $N_0(v, \cdot)$: number of unseen bigrams starting with v; $N_0(\cdot)$: number of unseen unigrams.

7.5.2 The full trigram model

To illustrate all the details involved in smoothing a trigram model, we will write down the full set of equations that is needed to smooth a trigram language model. Any of the three smoothing methods can be used for this purpose. Here, we will consider only absolute discounting in connection with interpolation. At the levels of bigrams and unigrams, we have to apply the same operation. Without using the singleton distribution, we obtain the full trigram model:

$$p_{tri}(w|uv) = \max\left\{0, \frac{N(u, v, w) - b_{uv}^{tri}}{N(u, v, \cdot)}\right\} + b_{uv}^{tri} \cdot \frac{W - N_0(u, v, \cdot)}{N(u, v, \cdot)} \cdot p_{bi}(w|v)$$

$$p_{bi}(w|v) = \max\left\{0, \frac{N(v, w) - b_v^{bi}}{N(v, \cdot)}\right\} + b_v^{bi} \cdot \frac{W - N_0(v, \cdot)}{N(v, \cdot)} \cdot p_{uni}(w|v)$$

$$p_{uni}(w) = \max\left\{0, \frac{N(w) - b^{uni}}{N(\cdot)}\right\} + b^{uni} \cdot \frac{W - N_0(\cdot)}{N(\cdot)} \cdot \frac{1}{W}$$
(7.20)

The advantage of interpolation in comparison with backing-off is that the computationally costly renormalisation is avoided; the experiments show that the perplexities are virtually not affected. The smoothing at the level of unigrams will be required only in exceptional cases when the size of the training corpus is extremely small. In all other conditions, we will set: $b_{uni} \equiv 0$.

Note that we have not used the singleton distribution. In this case, the normal bigram model has to be replaced by the corresponding singleton counts, that is

$$\frac{N_1(\cdot, v, w)}{N_1(\cdot, v, \cdot)} \qquad \text{rather than} \qquad \frac{N(v, w)}{N(v, \cdot)} \tag{7.21}$$

When using the singleton distribution, we have to be careful in using the leaving-one-out concept because the set of events now is not any more the full set of training data but only the corresponding singleton subset. As to the choice of history dependent versus history independent discounting parameters, it is important to know that the experimental tests show that the history dependence does not pay off, at least for absolute discounting. So it is sufficient to have only two (!) discounting parameters, namely one for the bigram model and another one for the trigram model.

7.5.3 Practical issues

Pooling of parameters

In smoothing trigram models by linear discounting or interpolation, the smoothing parameters, both λ_{uv} and b_{uv} , have been assumed to depend on the word pair uv. The number of these word pairs itself is huge so that reducing the number of smoothing parameters is desirable. Often, the parameters are *pooled or tied* across different histories by setting (Jelinek and Mercer 1980):

$$\lambda_h = \lambda_r \quad \text{for all } h \text{ with } \quad N(h) = r \tag{7.22}$$

This means that the parameters are tied across histories h with the same count N(h). It is straightforward to repeat the derivations for this type of tying. As result, we obtain:

$$\lambda_r = \frac{\sum\limits_{h:N(h)=r} N_1(h,\cdot)}{\sum\limits_{h:N(h)=r} N(h,\cdot)}$$
(7.23)

Similarly, when assuming the parameters to be *independent* of the histories h, we obtain:

$$\lambda = \frac{\sum\limits_{h}^{N} N(h, \cdot)}{\sum\limits_{h}^{N} N(h, \cdot)} = \frac{N_1(\cdot, \cdot)}{N(\cdot, \cdot)}$$
(7.24)

For absolute discounting, we obtain similar formulae in the case of tying. In particular for absolute discounting, the experimental results show that there is no degradation in perplexity when using history *independent* discounting parameters (Ney and Essen 1993).

Computational complexity

In a real implementation, no matter whether for off-line purposes or in an operational prototype system, the computational complexity of the trigram or bigram model has to be considered, namely the memory requirements and the access time.

- To store the trigram and bigram counts, we cannot use a standard array implementation. Even for a simple model like a bigram model, it is impossible to store the full table of conditional probabilities, which would require $20000^2 = 400 \cdot 10^6$ entries for a vocabulary of 20000 words. Instead, often a so-called list implementation is used which works as follows. For each word w, we have a pointer into the actual list, and for each observed trigram (u, v, w), an entry of the list consists of two parts, namely the index pair (u, v) and the trigram count N(u, v, w). For efficiency, this organisation should be combined with the counts for the bigram models. A further reduction in memory costs is achieved by removing the singletons (Katz 1987) from the list of seen bigrams and trigrams. In particular, by simply removing the trigram singletons, the storage costs can be reduced dramatically without affecting the perplexity much. We give a specific example for the WSJ corpus of 39 million running words, for which experimental results will be described. Here, the storage cost is 126 Mbyte when 4-byte values are used for all pointers and counts. This storage cost can be reduced down to 66 Mbyte by omitting the trigram singletons.
- For fast access to the bigram and trigram lists, we have to use suitable techniques, e.g. binary search. Short term storing of language probabilities might also be appropriate, depending on the interaction between language model and search.

7.5.4 Cache

The so-called cache model has been used successfully by a number of researchers (Kuhn and De Mori 1990; Jelinek et al. 1991b; Rosenfeld 1994). The cache can be viewed as a short-term memory where the probability of the most recent words is increased. In other words, the cache model takes into account that the words of the vocabulary are not distributed homogeneously over a text, but tend to occur in clusters. The typical mathematical formulation for the cache contribution is as follows:

$$p_{cache}(w_n | w_{n-M} \dots w_{n-1}) = \frac{1}{M} \sum_{m=1}^M \delta(w_{n-m}, w_n)$$
(7.25)

where $\delta(x, y)$ denotes the Kronecker function, which is 1 if the two arguments are the same and 0 otherwise. The probability of the cache model is typically combined with the trigram model by linear interpolation. There are refinements that suggest themselves:

- We can introduce weights that depend on the distance in terms of word positions; typically these weights should go smoothly to zero to introduce some sort of *forgetting*.
- In a number of cases such as dictation of documents, the beginning of a new document is known in most cases, and the cache should be reset to zero at the document boundaries.
- One can argue that the cache is most important for low frequency words and therefore should be used only for this subset of words.

The cache concept considered so far is based on unigrams only. As in the case of unigrams, we can argue that word bigrams and trigrams tend to occur in clusters, too. Extensions of the unigram cache to bigrams and/or trigrams have been successfully used in Jelinek et al. (1991b) and Rosenfeld (1994). For example, in the case of a bigram cache, the bigram counts based on the most recent history are used to compute the probabilities for the bigram cache. The cache model described here can be interpreted as a special case of so-called *adaptive* language models that adapt their probabilities to the most recent history, say the last 100 to 1000 predecessor words. In contrast, a *non-adaptive* language model does not depend on the test data, but remains unchanged as trained on the training data. For other types of adaptive language models see Essen and Steinbiss (1992) and Rosenfeld (1994).

7.5.5 Experimental results

To illustrate some of the issues in language modelling, we discuss some experimental results (Rosenfeld 1994; Generet et al. 1995). The results were obtained for a subset of the Wall Street Journal (WSJ) corpus. The vocabulary consisted of the (approximately) 20000 most frequent words. In addition, there were two non-speech words. First, each out-of-vocabulary word was replaced by a symbol for *unknown word*. Second, to mark the sentence end, a symbol for sentence boundary was added. There were three different training sets with 1, 5 and 39 million words and a separate test set of 0.325 million words as shown in Table 7.2.

Table 7.3 summarises some numbers from which we can estimate the coverage for the three training sets. This table gives the number of different bigrams and the number of singleton bigrams in training. The same numbers are also given for trigrams. As mentioned in the context of linear discounting, we can use these numbers to estimate the probability for new unseen trigrams. We obtain a probability of 0.57, 0.44 and 0.28 for the training sets of 1, 5 and 39 million words, respectively.

Table 7.2: Number of words and sentences in training and test data (vocabulary: about 20000 words)

	words	sentences
train-1	972 868	41 156
train-5	$4\ 513\ 716$	$189\ 678$
train-39	$38\;532\;517$	$1 \ 611 \ 572$
test	324 655	$13 \ 542$

Table 7.3: Number of distinct and of singleton events for bigrams and trigrams

	distinct bigrams	bigram singletons
train–1	303 858	211 105
train-5	$881\ 263$	$566 \ 093$
train-39	3 500 636	$2 \ 046 \ 462$
	distinct trigrams	trigram singletons
train-1	distinct trigrams 648 482	trigram singletons 556 185
train–1 train–5	distinct trigrams 648 482 2 420 168	trigram singletons 556 185 1 990 507

The perplexities for different conditions are summarised in Table 7.4. Table 7.4 consists of two parts, namely A) and B), for which the results are reported in Generet et al. (1995) and Rosenfeld (1994), respectively. In either part, for each language model test, there are three perplexities, namely for the three training sets of 1, 5 and 39 million words, so that the influence of the size of the training set on the perplexity can be seen immediately. Unfortunately, a direct comparison of the perplexities reported for the two parts of Table 7.4 is difficult for two reasons. First, due to small differences in selecting the articles from the Wall Street Journal corpus, the corpora used in the two parts of Table 7.4 are not completely identical. Second, the unknown word may be handled differently in the perplexity calculations.

For part A) of Table 7.4, the methods have been described in this chapter. The baseline method was absolute discounting with interpolation; the discounting parameters were history independent. The baseline trigram model was combined with extensions like the singleton backing-off distribution, and the cache model, which was tested in two variants, namely at the unigram level and at the combined unigram/bigram level. For comparison purposes, the baseline trigram language model was also compared with a

Size of training corpus	1 Mio	5 Mio	39 Mio		
A) absolute discounting and interpolation	ı (Gene	eret et al	. 1995)		
bigram model (with singletons)	288	217	168		
trigram model	250	167	105		
+ singleton	222	150	97		
+ unigram cache	191	133	90		
+ bi-/unigram cache	181	128	87		
+ singleton $+$ bi-/unigram cache	173	124	85		
B) Katz' backing-off (Rosenfeld 1994)					
trigram model	269	173	105		
+ bi-/unigram cache	193	133	88		
+ bi-/unigram cache + maximum entropy	163	108	71		

Table 7.4: Perplexities for different language models

bigram language model. Considering part A) of Table 7.4, we can see:

- The improvement obtained by a trigram model over a bigram model is as follows (for singleton backing-off): the perplexity goes down from 168 to 97 for the 39-million training corpus whereas for the 1-million corpus the improvement in perplexity is only from 288 to 222. So in both cases there is an absolute improvement of the perplexity of about 60.
- The singleton backing-off distributions help to improve the perplexity; the improvement is the higher the smaller the size of the training corpus is.
- In all tests, the cache model results in a significant improvement, and the addition of the bigram cache always results in an improvement over the pure unigram cache.
- The lowest perplexities were obtained for a combination of singleton distributions and the unigram/bigram cache. There is a significant improvement by just increasing the amount of training data: the perplexity is halved by increasing the training corpus from 1 million to 39 million words.

For the CMU results obtained by Rosenfeld (1994), the trigram model was based on Katz' backing-off (Katz 1987), which uses the Turing-Good formula (Good 1953). To reduce the memory requirements, the trigram singletons in the training data were omitted. The trigram model was combined with the two cache variants (unigram cache and bigram cache) and the maximum entropy model by linear interpolation.

When looking at the perplexities in part B) of Table 7.4, we see that they are better than those shown in part A) only in the case that the maximum entropy model is added. The maximum entropy model has two characteristic aspects (Rosenfeld 1994):
- The training is based on a criterion other than maximum likelihood, namely on the so-called maximum entropy principle.
- In addition to the usual bigram and trigram dependencies, the language model is designed to model distant bigrams and trigrams and word trigger pairs. A distant *m*-gram model allows for an additional insertion before the word under consideration. For example, the word tuple (v, x, w) with an arbitrary word x defines a distant bigram (v, *, w). Word trigger pairs are used to capture the dependence between selected word pairs over a longer distance (Rosenfeld 1994). Examples of such trigger word pairs are government parliament and Europe Brussels.

In comparison with the conventional trigram language model, the maximum entropy requires a much higher cost in terms of programming and CPU time for training. There is an improvement in the order of 20 %. This observation is in agreement with the general experimental experience with other language model techniques: To achieve an improvement over a baseline model like the trigram model in combination with the cache, a lot of effort is required, and even then the improvement may be small. For more details on word triggers and maximum entropy, see Bahl et al. (1984), Rosenfeld (1994), Lau et al. (1993), respectively.

To study the dependence of perplexity on the discounting parameters, experimental tests were carried out by Generet et al. (1995). Figure 7.3 and Figure 7.4 show the perplexity as a function of the discounting parameters. In all cases, the training was based on the 5-million corpus. For both figures, there were three conditions under which the perplexity was measured:

- training data without leaving-one-out: Obviously, this condition results in a much too optimistic estimate of the perplexity as can be seen from the comparison with the perplexities of the other two conditions.
- training data with leaving-one-out (LOO): This is the typical training condition, which we have described in this chapter. The perplexities are much higher than without leaving-one-out and already close to the perplexities of the test data.
- test data: This condition is the default condition for measuring a corpus perplexity, i.e. the test data are not used to train the language model. It can be seen that the perplexities are somewhat higher than the leaving-oneout perplexities. But still, the difference is not too big, which demonstrates the viability of the leaving-one-out concept.

In other words, the last two conditions correspond to the cross-validation concept: either we create a test set from the training data by leaving-one-out or we are given a completely separate set of test data.

When comparing the two types of smoothing, we can see that the perplexity curve for absolute discounting has a very flat minimum for the two crossvalidation conditions. This shows that the choice of the parameter b for absolute discounting is not critical. The perplexity curve for linear discounting shows a different behaviour in that the minimum is more distinct.



Figure 7.3: Perplexity as a function of b for absolute discounting with backing-off

The optimal perplexity for linear discounting is significantly higher than the optimal perplexity for absolute discounting. However, we have to remember that the linear discounting model here is based on history *independent* discounting parameters, and it is a well known fact that history dependence is important in the case of linear discounting (Jelinek and Mercer 1980).

7.5.6 Recommendations: *m*-gram language models

Here, we give recommendations on how to use language models for a specific application in speech recognition:

- 1. Today, by far the most successful method in language modelling is the bigram and trigram approach. If there are enough training data, a trigram model should be used; otherwise a bigram model might be sufficient.
- 2. Smoothing of the models is *always* required. When smoothing a trigram model with a bigram model, or a bigram model with a unigram model, our experience is that the method of absolute discounting is the method of choice: it is simple and robust with respect to the choice of the smoothing parameters. The backing-off method introduced by Katz (1987) produces comparable results at a somewhat higher effort.
- 3. The use of improved backing-off distributions like the singleton distribution plays only a minor role.



Figure 7.4: Perplexity as a function of λ for linear discounting with backing-off

- 4. In any application, it should be checked whether the cache effect applies. Examples of such applications are text dictation and maybe dialogue systems. In these cases, the cache model should be combined with the baseline trigram model.
- 5. When combining language models from "different sources", linear interpolation is the method of choice. Only in rare cases will it be necessary to go through the trouble of performing a full training with the EM algorithm. Even then it will be necessary in most cases to reduce the total number of independent interpolation parameters by tying.

7.6 Refined language models

7.6.1 Equivalence histories and word classes

In this section, we will consider structures that allow us to define classes (or categories) for words that exhibit similar behaviour in the context of language modelling. Word classes or parts of speech (POS) can be viewed as an attempt to cope with the problem of sparse data in language modelling (Derouault and Merialdo 1986). Typically, these word classes are based on syntactic-semantic concepts and are defined by linguistic experts. Generalising the concept of word similarities, we can also define word classes by

using a statistical criterion, which in most cases is, but does not have to be, maximum likelihood (Jelinek et al. 1990; Jelinek 1991; Brown et al. 1992; Ney et al. 1994).

There are two levels at which we will introduce equivalence concepts: the level of word histories h and the level of single words w. In a sequence of training data $w_1...w_n...w_N$, the history of word w_n is given by the most recent M words $w_{n-M}...w_{n-1}$. The equivalence classes of the possible word histories h for a word w will be called states and denoted by a so-called

state mapping $S: h \to s = S(h)$.

For the words w, we will use the so-called

class (or category) mapping $G: w \to g = G(w)$.

For both the states s and the word classes g, we define new counts:

$$N(s,w) := \sum_{h:S(h)=s} N(h,w)$$

$$N(s) := \sum_{w} N(s,w)$$

$$N(h,g) := \sum_{w:G(w)=g} N(h,w)$$

$$N(g) := \sum_{w:G(w)=g} N(w)$$
(7.26)

Using only the history mapping $h \to S(h)$, we have the probability model:

$$Pr(w|h) = p(w|S(h)).$$
 (7.27)

with the log-likelihood function:

$$F(S) = \sum_{h} N(h, w) \log p(w|S(h)) = \sum_{(s,w)} N(s, w) \log p(w|s)$$
(7.28)

Here, as usual, the symbol p(.) denotes a more specific model as opposed to Pr(.). Plugging in the relative frequencies as estimates for p(w|s) = N(s,w)/N(s), we obtain:

$$F(S) = \sum_{(s,w)} N(s,w) \log \frac{N(s,w)}{N(s)}$$
(7.29)

This criterion has the form of an entropy, which also arises in the context of hierarchical equivalence classes and CART (Bahl et al. 1989; Breiman et al. 1984), where a tree structure is imposed on the mapping S. In contrast, here no hierarchical structure is imposed on the word classes so that no

special structure for the mapping $h \to S(h)$ is assumed. By using equivalence states, we can reduce the number of free parameters: $|S| \cdot (W-1)$ in comparison with $W \cdot (W-1)$ for a bigram model and $W^2 \cdot (W-1)$ for a trigram model, where |S| is the number of equivalence states. However, these numbers are somewhat artificial, because even for large text databases, the number of really independent parameters is much smaller.

When now adding word classes using the mapping $w \to G(w)$, we have to distinguish two types of probability distribution:

- p(g|s): Probability that, given state s = S(h) with history h, class g is observed in the *next* position.
- $p_0(w|g)$: Probability that, given class g in the current position, word w is observed in the *current* position. For each word w, there is *exactly one* class G(w):

$$p_{0}(w|g) = \begin{cases} p_{0}(w|g) > 0 & \text{if } g = G(w) \\ 0 & \text{if } g \neq G(w) \end{cases}$$
(7.30)

Therefore, we can use the somewhat sloppy notation $p_0(w|G(w))$.

For the probability Pr(w|h), we then have the decomposition:

$$Pr(w|h) = p_0(w|G(w)) \cdot p(G(w)|S(h))$$
(7.31)

The log-likelihood function now depends on both mappings G and S and can be written as (Ney et al. 1994):

$$F(S,G) = \sum_{w} N(w) \log p_0(w) + \sum_{(s,g)} N(s,g) \log \frac{p(s,g)}{p(s)p(g)}$$
(7.32)

Plugging in the relative frequencies for p(w), p(s, g), p(s), p(g) as maximum likelihood estimates, we have:

$$F(S,G) = \sum_{w} N(w) \log N(w) + \sum_{(s,g)} N(s,g) \log \frac{N(s,g)}{N(s)N(g)}$$
(7.33)

A two-sided symmetric model can be defined by using the word classes both for the current word and its predecessor words. For a bigram (v, w), we use (g_v, g_w) to denote the corresponding word classes. For a bigram model, we have then:

$$Pr(w|v) = p_0(w|G(w)) \cdot p(G(w)|G(v))$$
(7.34)

The log-likelihood function for such a symmetric model is easily derived using the equations of the preceding paragraph:

Table 7.5:	Algorithm	for	word	clustering
------------	-----------	-----	------	------------

Star	Start with some initial mapping $G: w \to G(w)$		
Itera	Iterate until some convergence criterion is met		
	Loop	o over all words w	
		Loop over all clusters g'	
		Compute likelihood if w is moved from cluster	
		g = G(w) to g'	
		Move word w to the cluster g' with the highest	
		likelihood	

$$F_{bi}(G) = \sum_{w} N(w) \log N(w) + \sum_{(g_v, g_w)} N(g_v, g_w) \log \frac{N(g_v, g_w)}{N(g_v)N(g_w)}$$
(7.35)

where (g_v, g_w) denotes the class bigram of the word bigram (v, w) and $N(g_v, g_v), N(g_w), N(g_v)$ are defined in the usual way. Such a symmetric model leads to a drastic reduction in the number of free parameters: $|G| \cdot (|G|-1)$ probabilities for the table $p(g_w|g_v)$; (W-|G|) probabilities for the table $p_0(w|g_w)$, and W indices for the mapping $w \to G(w)$. In a similar way, we have for the symmetric trigram model:

$$F_{tri}(G) = \sum_{w} N(w) \log N(w) + \sum_{(g_u, g_v, g_w)} N(g_u, g_v, g_w) \log \frac{N(g_u, g_v, g_w)}{N(g_u, g_v)N(g_w)}$$
(7.36)

So far, the assumption has been that the mappings for the word classes or the equivalence states are known. Now we describe a procedure by which such mappings can be determined automatically. This automatic procedure will be developed for the two-sided symmetric model of Eq.(7.35). The task is to find a mapping $G: w \to g = G(w)$ that assigns each word to one of |G|different word classes. Since these classes are found by a statistical clustering procedure, they are also referred to as word clusters. The perplexity of the training data, i.e. $F_{bi}(G)$, is used as optimisation criterion. In the spirit of decision-directed learning (Duda and Hart 1973, p. 210), the basic concept of the algorithm is to improve the value of the optimisation criterion by making local optimisations, which means moving a word from one class to another in order to improve the log-likelihood criterion. Thus we obtain the algorithm shown in Table 7.5.

More details on this algorithm and related ones can be found in Brown et al. (1992) and Ney et al. (1994). For small amounts of training data, such automatically determined word classes may help to reduce the perplexity of a word based bigram or trigram language model (Brown et al. 1992; Ney et al. 1994).

7.6.2 Grammar based language models

Due to the smoothing techniques, bigram and trigram language models are robust and have been successfully used more widely in speech recognition than conventional grammars like context free or even context sensitive grammars. Although these grammars are expected to better capture the inherent structures of the language, they have a couple of problems:

- *robustness*: the grammar must be able to handle a vocabulary of 10000 or more words, and ultimately a non-zero probability must be assigned to each possible word sequence.
- *ambiguity*: while *m*-gram language models avoid any ambiguity in parsing, context free grammars are typically ambiguous and thus produce more than a single parse tree.

When using context free grammars, the terminals or symbols that can be observed are the words of the vocabulary (Lafferty et al. 1992; Wright et al. 1993; Della Pietra et al. 1994; Yamron 1994). We use the notation:

- A, B, C, ... non-terminals or abstract syntactic categories;
- u, v, w, ... terminals, i.e. the spoken words $w_1 ... w_n ... w_N$;
- α : an arbitrary nonempty string of terminals and non-terminals.

We typically attach probabilities to the context free rules. For a production rule like

$A \rightarrow \alpha$

we have a conditional probability $p(A \to \alpha | A)$ that a given non-terminal A will be rewritten as α . As usual, these probabilities must be normalised:

$$\sum_{\alpha} p(A \to \alpha | A) = 1$$

As usual in the framework of context free grammars, one often uses socalled normal forms like Chomsky and Greibach normal form for which standard introductions to formal languages in compiler construction theory can be consulted. Typically, these normal forms simplify the analysis of the problem under consideration, and there are automatic techniques for converting arbitrary context free grammars into these normal forms.

For each parse tree of a given string, we can compute its probability as the product over all its production rules. The probability of the given word string is then obtained by summing these probabilities over all possible parse trees. The fundamental definitions for context free grammars are well covered in textbooks (Fu 1982; Gonzalez and Thomason 1978).

For the application of stochastic context free grammars, we consider four questions or problems:

- parsing problem: How to find the best parse tree along with its probability? How to efficiently calculate the total probability of all parse trees for a given sentence? The solution is given by the stochastic extension of the CYK algorithm (Fu 1982; Gonzalez and Thomason 1978), which is sometimes referred to as *inside algorithm* (Jelinek et al. 1992).
- predicting the next word: How can we use a grammar to predict the next word, i.e. given the word sequence $w_1...w_{n-1}$, how can we compute the conditional probability for the word w_n in position n:

 $Pr(w = w_n | w_1 ... w_{n-1})$?

The solution is provided by the so-called *left corner algorithm* by Jelinek and Lafferty (Jelinek et al. 1992).

- *learning the rule probabilities*: Given the rules of a grammar, how can we estimate the rule probabilities from training sentences? The solution is provided by the corresponding version of the EM algorithm, which is called the *inside-outside algorithm* (Jelinek et al. 1992).
- grammatical inference: How can we learn the grammar as a whole from a set of training data (Fu 1982; Gonzalez and Thomason 1978)? It is this problem for which no good solutions exist yet.

The main stumbling block in using grammar based language models so far seems to be that the grammar rules as such are not available, either because handwritten grammars are simply not good enough or because the problem of grammatical inference is too big. To mitigate these problems, there have been attempts to use special types of context free grammars which are referred to as lexicalised or link grammars. Their non-terminals are assumed to depend on the words of the lexicon. These grammars include the bigram and trigram language models as special cases and thus provide a smooth transition from the m-gram language models to context free grammars (Lafferty et al. 1992; Della Pietra et al. 1994; Yamron 1994). Although so far there are only preliminary results, we will consider these approaches in more detail because they offer promising extensions from m-gram models and they provide a good framework to show the relationship between m-gram language models, finite state grammars and context free grammars.

To describe this type of grammar, we will use a special version of the Greibach normal form. Each context free grammar can be converted into a grammar whose rules are of the following three types:

$$\begin{array}{rccc} A & \rightarrow & w \\ A & \rightarrow & wB \\ A & \rightarrow & wBC \end{array}$$

The third rule type describes a branching process, which is required for the full power of context free grammars. Without it, we are restricted to regular



Figure 7.5: Bigram as a finite state grammar



Figure 7.6: Trigram as a finite state grammar

grammars which can be conveniently represented by finite state networks. Each network state is identified with a non-terminal of the grammar. To obtain a bigram language model in this framework, we use a separate non-terminal A_w for each word w. The bigram model itself is expressed by the following rule type for the word pair (vw):

 $A_v \rightarrow wA_w$

The bigram rules can be represented as a finite state network as illustrated in Figure 7.5 for a given word sequence. Unlike the conventional representation, the nodes stand for the observed words, i.e. the terminals, whereas the links represent the non-terminals. For a complete description of the grammar, we have to include the rules for sentence begin and end. For a trigram model, we have to use a separate non-terminal for each word pair (vw), and the rules depend on the word triples (uvw):

 $A_{uv} \rightarrow wA_{vw}$

The trigram rules are illustrated in Figure 7.6.

To go beyond the power of *m*-gram and finite state language models, we need more than one non-terminal on the right-hand side. To simplify the presentation, we restrict ourselves to the special case of pairwise word dependencies, which can be viewed as an extended bigram model. In addition to the standard bigram non-terminals now denoted by L_w , we have additional non-terminals for the branching process, which are denoted by R_w for word w:

$L_v \rightarrow$	w	$(halt\ rule)$
$L_v \rightarrow$	wL_w	$(step\ rule)$
$L_v \rightarrow$	wL_wR_w	(branch rule)

The production rule for non-terminal R_w is different from the rules for L_w and, for example, could have this form:

$$R_w \rightarrow yL_y$$



Figure 7.7: Parsing in the conventional form

This type of rule is capable of handling long range dependencies¹ sich as those in the following English examples (Della Pietra et al. 1994):

between ... and ... neither ... nor ... to describe ... as ... to prevent ... from ...

A conventional bigram or trigram language model cannot capture these dependencies. In general, for a given word string, context free grammars define a hierarchical structure by the parsing process. Figure 7.7 shows the parsing structure in the conventional form. There is another equivalent representation shown in Figure 7.8, where the long range dependencies are expressed by special *links* in addition to the short range (i.e. next neighbour) bigram links. In this link representation, the three different types of rule can be seen as follows. The step rule stands for the standard bigram link to the next right word. The branch rule has two links: in addition to the bigram link, there is a long range link to another word, e.g. the link from word w_3 to w_7 in Figure 7.8. Since there is the long range link for word w_7 , there is no bigram link for w_7 , and therefore the *halt* rule must be used for word w_6 , the predecessor of w_7 . For this type of grammar based language model, there are only some preliminary experimental results (Della Pietra et al. 1994; Yamron 1994). A still existing problem is the question of how to select the relevant word links in view of the large number of possible spurious links. In Della Pietra et al. (1994), a method has been suggested for inferring the links automatically from a text corpus.

¹In linguistics, *long range dependency* refers rather to the relation between an antecedent pronominal item and the position where this item satisfies verb valency conditions, as in *Where did John tell Mary he thought Ted was convinced Henry had put it* EMPTY ADVERBIAL POSITION ? It is known that these can also be handled by context-free mechanisms. (Editor's note.)



Figure 7.8: Parsing in link grammar form

7.6.3 Recommendations: Refined language models

Here, some recommendations are given for the use of the refined language models in specific recognition tasks:

- 1. Experimental experience is that any type of the usual language model refinements is unlikely to reduce the perplexity by more than 10% over a standard trigram model (or bigram model, if the amount of training data is small). Therefore in all applications, it should be checked first whether a trigram model in combination with a cache component does not already do the job. In a number of recognition tasks, the perplexity improvements by the language model refinements are not worth the additional effort using today's algorithms.
- 2. There might be some particular applications where the amount of training data is really small. In these cases, it can be useful to base the language model on word classes rather than the words themselves. These word classes can be classes defined either by an automatic clustering procedure or by linguistic prior knowledge, e.g. parts of speech (POS).
- 3. If it is suitable to combine two language models of different type, e.g. a word bigram model and a class bigram model, the first choice should be to try a linear interpolation of the two models.

7.7 Language models and search

7.7.1 Sources of recognition errors

Looking at the basic architecture shown in Figure 7.1 we see that there are different types of reason why a speech recognition system, in particular a large-vocabulary continuous-speech system, can make a recognition error:

- *acoustic-phonetic modelling*: This part of the system includes all parts related to the acoustic signal:
 - signal analysis;
 - phoneme modelling:
 - the inventory of context independent and context dependent phoneme units;

- in most cases, the phoneme units are represented by Hidden Markov models (Levinson et al. 1983; Bahl et al. 1983); any of their details such as topology and emission probabilities may have an effect on the error rate;
- pronunciation lexicon: the pronunciation lexicon serves as the link between the word level and the phoneme units.

It is obvious that any of these three levels of acoustic-phonetic modelling can cause recognition errors. For example, a word whose entry in the pronunciation lexicon is incorrect is unlikely to be recognised correctly.

- *language modelling*: If the language model is poor it cannot help much to resolve the ambiguities in acoustic recognition.
- search errors: A full, i.e. globally optimal, search is prohibitive for large vocabulary speech recognition. Therefore global optimal search is abandoned and replaced by a suboptimal search. Not finding the globally optimal word sequence may cause additional recognition errors. These search errors will disappear if the search effort is increased to evaluate more hypotheses about the spoken word sequences.

7.7.2 Search: Finding the single best sentence

In a suboptimal search, which must be used in a real recognition system, there is no guarantee to find the global optimum as required by the Bayes decision rule. Therefore it is important to organise the search as efficiently as possible to minimise the number of search errors. Here, we will describe so-called time synchronous beam search which is the key component in many large vocabulary systems (Alleva et al. 1993; Fissore et al. 1993; Murveit et al. 1993; Soong and Huang 1991; Haeb-Umbach and Ney 1994; Valtech et al. 1994).

Acoustic-phonetic modelling is based on Hidden Markov models. For an utterance to be recognised, there is a huge number of possible state sequences, and all combinations of state and time must be systematically considered. In the so-called Viterbi or maximum approximation (Bahl et al. 1983; Levinson et al. 1983), the sum over all paths is approximated by the path which has the maximum contribution to the sum:

$$[w_1^N]_{opt} = \arg \max_{[w_1^N]} \left\{ Pr(w_1^N) \cdot \sum_{[s_1^T]} Pr\left(x_1^T; s_1^T \middle| w_1^N\right) \right\}$$
(7.37)

$$\cong \arg \max_{[w_1^N]} \left\{ Pr(w_1^N) \cdot \max_{[s_1^T]} Pr\left(x_1^T; s_1^T \middle| w_1^N\right) \right\}$$
(7.38)

where the sum is to be taken over all paths $[s_1^T]$ that are consistent with the word sequence $[w_1^N]$ under consideration.

In the maximum approximation, the search problem can be specified as follows. We wish to assign each acoustic vector at time t to a (state,word)



Figure 7.9: Example of a time alignment path

index pair. This mapping can be viewed as a time alignment path, which is a sequence of (state,word) index pairs (stretching notation):

$$(s_1, w_1), \ldots, (s_t, w_t), \ldots, (s_T, w_T)$$

An example of such a time alignment path in connected word recognition is depicted in Figure 7.9. The dynamic programming strategy to be presented will allow us to compute the probabilities

$$Pr(w_1...w_N) \cdot Pr(x_1...x_T; s_1...s_T | w_1...w_N)$$

in a left-to-right fashion over time t and to carry out the optimisation over the unknown word sequence at the same time. Within the framework of the maximum approximation, the dynamic programming algorithm presents a closed-form solution for handling the interdependence of non-linear time alignment, word boundary detection and word identification in continuous speech recognition. Taking this strong interdependence into account has been the basis for many different variants of dynamic programming solutions to the problem (Vintsyuk 1971; Sakoe 1979; Bridle et al. 1982; Ney 1984). The sequence of acoustic vectors extracted from the input speech signal is processed strictly from left to right. The search procedure works



Figure 7.10: Bigram language model and search

with a time-synchronous breadth-first strategy, i.e. all hypotheses for word sequences are extended in parallel for each incoming acoustic vector. Originally, the dynamic programming strategy was designed for small vocabulary tasks like digit string recognition. To extend this algorithm towards large vocabulary tasks, there are two concepts that have to be added:

- pruning: Dynamic programming is still a full search (within the maximum approximation). By incorporating a pruning strategy, the full search is converted to a so-called beam search focussed on the most promising hypotheses (Lowerre and Reddy 1980). Since all hypotheses cover the same portion of the input, their scores can be directly compared. Every 10 ms, the score of the best hypothesis is determined, then all hypotheses whose scores fall short of this optimal score by more than a fixed factor are pruned, i.e. are removed from further consideration. For an efficient organisation, see Ney et al. (1992); for the details and refinements of the pruning strategy, see Steinbiss et al. (1994). The experimental tests indicate that for this type of beam search, depending on the acoustic input and the language model constraints, only a few percent of the potential state hypotheses have to be processed for every 10 ms of the input speech while at the same time the number of recognition errors is virtually not increased.
- *lexicon tree*: The experimental results show that the lion's share of the search effort is concentrated in the first two or three phonemes of each word. Therefore, to reduce the search effort at the word beginnings, it is important to organise the pronunciation lexicon in the form of a tree (Haeb-Umbach and Ney 1994).

Figure 7.10 illustrates how a bigram language model is incorporated into the search, where the vocabulary consists of the three words A, B, C. When

starting up a new word w, all predecessor words v have to be considered along with their bigram probabilities p(w|v). Note that, since the bigram language model is a special case, the set of word sequences shown in Figure 7.2 can be compressed into a finite state network.

When using a lexicon tree to represent the pronunciation lexicon, we are faced with the problem that the identity of the word is known only when we have reached the end of the tree. The solution is to use a separate lexicon tree in search for each predecessor word (Haeb-Umbach and Ney 1994) as illustrated in Figure 7.11.



Figure 7.11: Word recombination for a lexical tree

Fortunately, in the framework of beam search, the number of surviving predecessor words and thus the number of active trees is small, say 50 even for a 20000 word task with a perplexity of 200.

To prepare the ground for dynamic programming, we introduce two auxiliary quantities, namely the score $Q_v(t, s; w)$ and the back pointer $B_v(t, s; w)$:

- $Q_v(t, s; w) :=$ score of the best path up to time t that ends in state s of word w for predecessor v.
- $B_v(t,s;w) :=$ starting time of the best path up to time t that ends in state s of word w for predecessor v.

Note that for both quantities, the index w is known only at the leaves of the lexicon trees. Both quantities are evaluated using the dynamic programming recursion for $Q_v(t,s;w)$:

$$Q_{v}(t,s;w) = \max_{\sigma} \{ q(x_{t},s|\sigma;w) \cdot Q_{v}(t-1,\sigma;w) \}$$
(7.39)
$$B_{v}(t,s;w) = B_{v}(t-1,\sigma_{max}(t,s;w,v);w) ,$$

where $q(x_t, s | \sigma; w)$ is the product of emission and transition probability as given by the Hidden Markov model, and $\sigma_{max}(t, s; w, v)$ is the optimum predecessor state for the hypothesis (t, s; w) and predecessor word v. The back pointers $B_v(t, s; w)$ are propagated according to the dynamic programming (DP) decision. Unlike the predecessor word v, the index w is well defined only at the end of the lexical tree. Using a suitable initialisation for $\sigma = 0$, this equation includes optimisation over the unknown word boundaries. At word boundaries, we have to find the best predecessor word for each word. To this purpose, we define:

$$H(w;t) := \max \{ p(w|v) \cdot Q_v(t, S(w); w) \}$$
(7.40)

To start words, we have to pass on the score and the time index:

$$Q_v(t-1,0;w) = H(v;t-1)$$

$$B_v(t-1,0;w) = t-1$$
(7.41)

Note that here the right-hand side does not depend on word index w since as said before the identity of the spoken word is known only at the end of a tree.

To measure the search effort there are a number of quantities that can be used:

- The most important quantity is the average *number of state hypotheses* per 10-ms segment of speech. Strictly speaking, it should also be specified whether this number is measured *before* or *after* the pruning step.
- To study the distribution of the search effort over the search space in more detail, it can be useful to include the average number of *lexical trees*, of *word ends* and of *phoneme arcs*.
- For the dictation of text with a duration of several minutes, the storage requirements for bookkeeping can increase drastically. So a measure for the bookkeeping cost, both in terms of time and memory, should also be included.

7.7.3 Search: Word graphs

The main idea of a word graph is to generate word alternatives in regions of the speech signal, where ambiguity in the acoustic recognition is high.

The advantage is that pure acoustic recognition is decoupled from the application of the language model and that a complex language model, for example a stochastic context free grammar, can be applied in a subsequent postprocessing step. The number of word alternatives should be adapted to the level of ambiguity in acoustic recognition. The difficulty in efficiently constructing a good word graph is the following: the start time of a word depends in general on the preceding words. As a first approximation, we limit this dependence to the immediate predecessor word and obtain the so-called word pair approximation:

Given a word pair and its ending time, the word boundary between the two words is independent of the further predecessor words.

This word pair approximation was originally introduced in Schwartz and Austin (1991) to efficiently calculate multiple or *n*-best sentences. The word graph can be expected to be more efficient than the *n*-best approach. In the word graph approach, word hypotheses need to be generated only locally whereas in *n*-best methods each local alternative requires a whole sentence to be added to the *n*-best list. To give an oversimplified example, as shown in Figure 7.12, suppose we have 7 spoken words and 2 word hypotheses for each word position. The *n*-best method then requires $2^7=128$ sentence hypotheses, whereas the word graph approach produces a graph of only $2 \cdot 7 = 14$ word arcs.



Figure 7.12: Simplified example of a word graph

There have been a number of attempts at using two explicit levels in search: the first level produces a short list of either single word or sentence hypotheses, and at the second level, the final decision is taken using a complex language model (Sakoe 1979; Schwartz and Austin 1991; Soong and Huang 1991; Ney and Aubert 1994; Fissore et al. 1993; Oerder and Ney 1993; Woodland et al. 1995). In the following, we will describe a method for word graph generation based on word pair approximation. This method has been used successfully in two independent systems for 20000-word recognition (Ney and Aubert 1994; Woodland et al. 1995). The fundamental problem of word graph construction is:

- Given a word w and ending time t, how can we find a limited number of "most likely" predecessor words?
- This task is difficult since the start time of word w may very well depend on the predecessor word under consideration, which results in an interdependence of start times and predecessor words.

The basic idea is to isolate the probability contributions of a particular word hypothesis with respect to both the language model and the acoustic model. This decomposition can be visualised as follows:

$$\underbrace{x_1, \cdots, x_\tau}_{\text{words } w_1^{n-1}} \underbrace{x_{\tau+1}, \cdots, x_t}_{\text{word } w_n} \underbrace{x_{t+1}, \cdots, x_T}_{\cdots}$$

The set of likely word sequences is represented by a word graph, in which each arc represents a word hypothesis. Each word sequence contained in the word graph should be close (in terms of scoring) to the single best sentence produced by the one-pass algorithm. We prefer the term "word graph" to "word lattice" to indicate that when combining word hypotheses, we do not allow overlaps or gaps along the time axis. When using an *m*-gram language model $p(u_m | u_1^{m-1})$, we can recombine word sequence hypotheses at the phrase level if they do not differ in their final (*m*-1) words. Therefore it is sufficient to distinguish partial word sequence hypotheses only by their final words u_2^m . To describe the word graph construction algorithm, we introduce the following definitions:

- $h(w; \tau, t) := Pr(x_{\tau+1}^t | w) =$ probability that word w produces the acoustic vectors $x_{\tau+1}...x_t$.
- $H(u_2^m; t)$: (joint) probability of generating the acoustic vectors $x_1...x_t$ and a word sequence w_1^n with the final words $u_2^m = w_{n-m+2}$:

$$H(u_2^m;t) := \max_{w_1^n} \left\{ Pr(w_1^n) \cdot Pr(x_1^t | w_1^n) : w_{n-m+2}^n = u_2^m \right\}$$
(7.42)

where, as expressed by the notation, the final portion u_2^m of the word sequence w_1^n is not subjected to the maximisation operation.

Using the above definition, we can write the dynamic programming equation at the word level:

$$H(u_2^m;t) = \max_{u_1} \left\{ p(u_m | u_1^{m-1}) \ H(u_1^{m-1}; \tau(t; u_1^m)) \ h(u_m; \tau(t; u_1^m), t) \right\}$$
(7.43)

Here we have used the function $\tau(t; u_1^m)$ to denote the word boundary between u_{m-1} and u_m for the word sequence with final portion u_1^m and ending time t. To achieve a better pruning effect, a bigram language model is included. The word boundary itself is defined by a maximisation operation:

$$\tau(t; u_1^m) := \arg \max \left\{ H(u_1^{m-1}; \tau) \ h(u_m; \tau, t) \right\}$$
(7.44)

So far this has been just a notational scheme for the word boundary function $\tau(t; u_1^m)$. The crucial assumption now is that the dependence of the word boundary $\tau(t; u_1^m)$ can be confined to the final word pair u_{m-1}^m . The



Figure 7.13: Word pair approximation

justification is that the other words have virtually no effect on the position of the word boundary in the word pair (u_{m-1}, u_m) (Schwartz and Austin 1991). This so-called word pair approximation is illustrated in Figure 7.13. From this figure, it is obvious that the assumption of the word pair approximation is satisfied the better the longer the predecessor word u_{m-1} is: all time alignment paths then are recombined before they reach the final state of the predecessor word. We express this word pair approximation by the equation:

$$\tau(t; u_1^m) = const(u_1^{m-2}) \tag{7.45}$$

As long as only a bigram language model is used, the word pair approximation is still exact (within the beam search approximation).

To compute the word boundary function $\tau(t; v, w)$, we have to distinguish the hypotheses in dynamic programming according to the predecessor word, and we therefore use the predecessor conditioned algorithm. We add two equations, namely one for calculating the word boundary function $\tau(t; v, w)$ and one for calculating the word score $h(w; \tau, t)$. The word boundaries are obtained using the back pointers at the word ends:

$$\tau(t; v, w) = B_v(t, S(w); w)$$
(7.46)

For each predecessor word v along with word boundary $\tau = \tau(t; v, w)$, the word scores are recovered using the equation:

$$h(w;\tau,t) := \frac{Q_v(t, S(w); w)}{H(v;\tau)}$$
(7.47)

where we obtain H(w; t) according to its definition:

$$H(w;t) = \max \{ p(w|v) \cdot Q_v(t, S(w); w) \}$$
(7.48)

The details of the algorithm are summarised in Table 7.6. The operations are organised in two levels: the acoustic level and the word pair level. At the end of the utterance, the word graph is constructed by backtracking through the bookkeeping lists.

Table 7.6: One-pass algorithm for word graph construction ("single best" and "word graph")

proceed over time t from left to right			
	AC	OUSTIC LE	VEL: process states
		- initialisati	on: $Q_v(t-1,0;w) = H(v;t-1)$
		- time align	$\mathrm{ment}:\;Q_v(t,s;w)\;\mathrm{using}\;\mathrm{DP}$
		- propagate	back pointers in time alignment
		- prune unl	ikely hypotheses
		- purge boo	kkeeping lists
	W	RD PAIR L	EVEL: process word ends
		single best:	for each pair $(w; t)$ do
		H(w; t)	$dt = \max_v \left[p(w v) \; Q_v(t,S(w);w) ight]$
		$v_0(w;;$	$f(t) = rg\max_v \left[p(w v) \; Q_v(t, S(w); w) ight]$
		- store	e best predecessor $v_0 := v_0(w; t)$
		- store	e best boundary $ au_0 := au(t; v_0, w)$
		word graph	: for each triple $(t; v, w)$ store
		- word	boundary $ au(t;v,w)$
		- word	l score $h(w; au, t)$
trace back:		ick: single	best sentence
		and w	ord graph

What has to be added to the single-best one-pass strategy is bookkeeping at the word level: rather than just the best surviving hypothesis, the algorithm has to memorise all the word sequence hypotheses that are recombined into just one hypothesis to start up the next lexical tree (or word models). In the single-best method, only the surviving hypothesis (v_0, τ_0) has to be kept track of. To reduce the memory cost, a method for removing obsolete

proceed over time t from left to right				
	process each word pair u_{m-1}^m in the graph			
		get word boundaries $\tau(.)$ and scores $h(.)$		
		\mathbf{pr}	ocess all word sequences u_1^m	
			$\widehat{H}(u_1^m;t) := p(u_m u_1^{m-1}) H(u_1^{m-1};\tau) h(u_m;\tau,t)$	
			$H(u_2^m;t) = \max_{u_1} \widehat{H}(u_1^m;t)$	
			$B(u_2^m;t) = \arg\max_{u_1} \widehat{H}(u_1^m;t)$	
trace back: use back pointers $\{B(u_2^m;t)\}$				

Table 7.7: DP algorithm for search through word graph

hypotheses must be included along with list organisations and dynamic construction of the search space (Ney et al. 1992). Given the word graph and the language model, the final search can be carried out at the word level using a left-to-right dynamic programming algorithm as shown in Table 7.7. The cost of this search through the word graph depends on the complexity of the language model. For a trigram model, it is typically less than 1% of the effort for constructing the word graph.

An example of a word graph for a three-word vocabulary A, B, C including silence at the sentence beginning and end is shown in Figure 7.14 on page 280.



Figure 7.14: Example of a word graph

The arcs stand for word hypotheses, whereas the circles along with the word name denote the word end. In this figure, the word pair approximation is reflected by the fact that the nodes are labelled with the identity of the incoming words and thus the dependence on the predecessor word can be taken into account. In reality, the word graph is more complicated because we have to allow for optional silence between the words.

When we try to measure the efficiency of a word graph, it is evident that there are two important quantities to be considered, namely the size of the word graph and its quality in terms of errors. Obviously, the two quantities are dependent on each other. We try to specify how they could be measured:

- *size of the word graph*: We can measure the number of *arcs and nodes* and normalise them with respect to the number of actually spoken words to account for different sentence lengths.
- errors of the word graph: What we want to know is how well is the actually spoken word sequence represented by the word graph? An error is encountered when there is no path through the word graph for the spoken word sequence. Such an error can be related to a word substitution, a word deletion or a word insertion. To calculate the exact number of word errors, the so-called Levenshtein distance is extended to compute the best path for the spoken word sequence through the word graph, where we allow for word substitutions, word deletions and word insertions.

7.7.4 Recommendations: Language models and search

The recommendations given here apply more to research prototype systems than real life applications:

- 1. For tasks using small vocabularies as in voice commands, the use of N-best method is sufficient.
- 2. For more complex tasks with large vocabularies, word graphs offer many advantages. In particular, they allow the testing of different language models without the need to rerun a full acoustic recognition.
- 3. For speech recognition in the context of understanding and dialogue systems (see Chapter 13), there is an obvious need for word graphs. Without using word graphs, it would still be possible to do the recognition job, but by using word graphs, it is much simpler to achieve an efficient algorithm for the interaction between acoustic recognition, language modelling and understanding.

7.8 Final note: the mathematics of smoothing

In this appendix, we present some more mathematical details for the smoothing techniques presented above. This appendix has been written for readers who are interested in the derivations of the parameter estimation formulae. To give a self-contained presentation, it was necessary to repeat parts from the previous sections.

7.8.1 Linear discounting and backing-off

The model of linear discounting in conjunction with backing-off (Katz 1987; Jelinek 1991) has the advantage that it results in relatively simple formulae.

The model is:

$$p(w|h) = \begin{cases} (1 - \lambda_h) \cdot \frac{N(h, w)}{N(h, \cdot)} & \text{if } N(h, w) > 0\\ \lambda_h \cdot \left[\frac{\beta(w|\overline{h})}{\sum\limits_{w':N(h, w') = 0} \beta(w'|\overline{h})} \right] & \text{if } N(h, w) = 0 \end{cases}$$
(7.49)

Here we have two types of parameters to be estimated:

- the discounting parameters λ_h for each history h
- the backing-off distribution $\beta(w|\overline{h})$ for a generalised history \overline{h} . Note that for each history h the generalised history \overline{h} must be well defined in order to have a backing-off distribution $\beta(w|\overline{h})$.

The unknown parameters are estimated by maximum likelihood in combination with leaving-one-out. We obtain the log-likelihood function:

$$F(\{\lambda_{h}\},\{\beta(w|\overline{h})\}) = \sum_{n} \log p_{(-1)}(w_{n}|h_{n})$$
(7.50)
$$= \sum_{h} \sum_{w} N(h,w) \log p_{(-1)}(w|h)$$

$$= \sum_{h} \sum_{w:N(h,w)=1} 1 \cdot \log \left[\lambda_{h} \cdot \frac{\beta(w|\overline{h})}{\sum_{w':N(h,w')=0} \beta(w'|\overline{h})}\right]$$

$$+ \sum_{h} \sum_{w:N(h,w)>1} N(h,w) \cdot \log \left[(1-\lambda_{h}) \cdot \frac{N(h,w)-1}{N(h,\cdot)-1}\right]$$

where $p_{(-1)}(w|h)$ denotes the probability distribution for leaving out the event (h, w) from the training data.

By doing some elementary manipulations, we can decompose the loglikelihood function into two parts, one of which depends only on λ_h and the other depends only on $\beta(w|\overline{h})$:

$$F(\{\lambda_h\},\{\beta(w|\overline{h})\}) = F(\{\lambda_h\}) + F(\{\beta(w|\overline{h})\})$$

+ const($\{\lambda_h\},\{\beta(w|\overline{h})\})$ (7.51)

The λ_h dependent part is:

$$F(\{\lambda_h\}) = \sum_{h} \left[\sum_{w:N(h,w)=1} 1 \cdot \log \lambda_h + \sum_{w:N(h,w)>1} N(h,w) \cdot \log (1-\lambda_h)\right]$$
$$= \sum_{h} \left[N_1(h,\cdot) \log \lambda_h + [N(h,\cdot) - N_1(h,\cdot)] \log (1-\lambda_h)\right]$$
(7.52)

Taking the partial derivatives with respect to λ_h and equating them to zero, we obtain the closed-form solution:

$$\lambda_h = \frac{N_1(h, \cdot)}{N(h, \cdot)} \tag{7.53}$$

The same value is obtained when we compute the probability mass of unseen words in the training data for a given history h:

$$\sum_{w:N(h,w)=0} p(w|h) = \frac{N_1(h,\cdot)}{N(h,\cdot)}$$
(7.54)

To estimate the backing-off distribution $\beta(w|\overline{h})$, we rearrange the sums:

 $F(\{\beta(w|\overline{h})\})$

$$=\sum_{h}\sum_{w:N(h,w)=1}\log\left(\frac{\beta(w|\overline{h})}{\sum\limits_{w':N(h,w')=0}\beta(w'|\overline{h})}\right)$$
(7.55)
$$=\sum_{\overline{h}}\sum_{h\in\overline{h}}\sum_{w:N(h,w)=1}\log\left(\frac{\beta(w|\overline{h})}{\sum\limits_{w':N(h,w')=0}\beta(w'|\overline{h})}\right)$$
$$=\sum_{\overline{h}}\left[\sum_{w}N_{1}(\overline{h},w)\log\beta(w|\overline{h}) - \sum_{h\in\overline{h}}N_{1}(h,\cdot)\log\left(\sum_{w':N(h,w')=0}\beta(w'|\overline{h})\right)\right]$$

where $N_1(h, \cdot)$ is the number of singletons (h, w) for a given history h, i.e. the number of words following h exactly once, and where $N_1(\overline{h}, w)$ is defined as:

$$N_1(\overline{h}, w) = \sum_{h \in \overline{h}: N(h, w) = 1} 1$$
(7.56)

Taking the derivative, we have:

$$\frac{\partial F(\{\beta(w|\overline{h})\})}{\partial \beta(w|\overline{h})} = \frac{N_1(\overline{h},w)}{\beta(w|\overline{h})} - \sum_{h\in\overline{h}} \left[\frac{N_1(h,\cdot)}{\sum_{w':N(h,w')=0} \beta(w'|\overline{h})} \cdot \frac{\partial}{\partial \beta(w|\overline{h})} \sum_{w':N(h,w')=0} \beta(w'|\overline{h}) \right]$$
$$= \frac{N_1(\overline{h},w)}{\beta(w|\overline{h})} - \sum_{h\in\overline{h}:N(h,w)=0} \left[\frac{N_1(h,\cdot)}{\sum_{w':N(h,w')=0} \beta(w'|\overline{h})} \right] = 0, \qquad (7.57)$$

where we have taken into account that there are only contributions from those histories h which appear in the sum over w'. We do not know a closed-form solution for $\beta(w|\overline{h})$. By extending the sum over all histories h(Kneser and Ney 1995), we obtain the approximation:

$$\beta(w|\overline{h}) \cong \frac{N_1(\overline{h}, w)}{\sum\limits_{w'} N_1(\overline{h}, w')}$$
(7.58)

For convenience, we have chosen the normalisation $\sum_{w} \beta(w|\overline{h}) = 1$. This type of backing-off distribution will be referred to as singleton distribution.

7.8.2 Linear interpolation

In linear interpolation, a weighted average between the relative frequencies $N(h, w)/N(h, \cdot)$ and the general distribution $\beta(w|\overline{h})$ is computed:

$$p(w|h) = (1 - \lambda_h) \frac{N(h, w)}{N(h, \cdot)} + \lambda_h \beta(w|\overline{h}) \quad \text{with} \quad \sum_w \beta(w|\overline{h}) = 1$$
(7.59)

In leaving-one-out, we have to renormalise by the following substitution:

$$\frac{N(h,w)}{N(h,\cdot)} \rightarrow \frac{N(h,w)-1}{N(h,\cdot)-1} := q(w|h)$$

$$(7.60)$$

The leaving-one-out likelihood function is:

$$F(\{\lambda_h\};\{\beta(w|\overline{h})\}) = \sum_{hw} N(h,w) \log \left[(1-\lambda_h) q(w|h) + \lambda_h \beta(w|\overline{h}) \right].$$
(7.61)

Taking the partial derivatives and doing some term rearrangements using q(w|h) = 0 for N(h, w) = 1, we obtain the iteration formulae (Ney et al. 1994):

$$\bar{\lambda}_{h} = \frac{1}{N(h,\cdot)} \sum_{w} N(h,w) \frac{\lambda_{h} \beta(w|\overline{h})}{(1-\lambda_{h})q(w|h) + \lambda_{h} \beta(w|\overline{h})}$$

$$= \frac{1}{N(h,\cdot)} \left[N_{1}(h,\cdot) + \sum_{w:N(h,w)>1} N(h,w) \frac{\lambda_{h} \beta(w|\overline{h})}{(1-\lambda_{h})q(w|h) + \lambda_{h} \beta(w|\overline{h})} \right].$$
(7.62)

Similarly, we have for the $\beta(w|\overline{h})$:

$$\bar{\beta}(w|\bar{h}) = \frac{\sum_{h \in \bar{h}} \left[N(h,w) \frac{\lambda_h \beta(w|\bar{h})}{(1-\lambda_h)q(w|h) + \lambda_h \beta(w|\bar{h})} \right]}{\sum_{h \in \bar{h}} \sum_{w'} \left[N(h,w') \frac{\lambda_h \beta(w'|\bar{h})}{(1-\lambda_h)q(w'|h) + \lambda_h \beta(w'|\bar{h})} \right]}$$

$$= \frac{N_1(\bar{h},w) + \sum_{h \in \bar{h}:N(h,w)>1} \left[N(h,w) \frac{\lambda_h \beta(w|\bar{h})}{(1-\lambda_h)q(w|h) + \lambda_h \beta(w|\bar{h})} \right]}{N_1(\bar{h},\cdot) + \sum_{h \in \bar{h}} \sum_{w':N(h,w')>1} \left[N(h,w') \frac{\lambda_h \beta(w'|\bar{h})}{(1-\lambda_h)q(w'|h) + \lambda_h \beta(w'|\bar{h})} \right]}$$

$$(7.63)$$

Again, the dominating effect of the singletons is evident. For the special choice $\beta(w|\overline{h}) = 0$ if N(h, w) > 0, we have linear discounting again and recover the corresponding equations.

The model of linear interpolation considered so far is special because it involves only two types of histories, namely h and \overline{h} . In general, we may have several generalised histories \overline{h}_i for i = 1, 2, ... For this general case of linear interpolation, the so-called EM algorithm provides an efficient iterative procedure for estimating the unknown parameters (Baum 1972; Dempster et al. 1977; Jelinek and Mercer 1980). To consider the details of the EM algorithm, we have to specify the full interpolation model:

$$p(w|h) = \sum_{i} \lambda(i|h) \cdot \beta(w|i, \overline{h}_{i})$$

$$\sum_{i} \lambda(i|h) = 1$$

$$\sum_{w} \beta(w|i, \overline{h}_{i}) = 1$$
(7.64)

For the interpolation to work, the interpolation parameter $\lambda(i|h)$ must explicitly depend on the index *i* of the generalised history. The framework of the EM algorithm is based on the so-called $Q(\mu; \overline{\mu})$ function, where $\overline{\mu}$ is the new estimate obtained from the previous estimate μ (Baum 1972; Dempster et al. 1977). The parameter μ stands for the whole set of parameters to be estimated. The $Q(\mu; \overline{\mu})$ function is an extension of the usual log-likelihood function and is for our model:

$$Q(\cdot) = Q(\{\lambda(i|h)\}, \{\beta(w|i,\overline{h}_i)\}; \{\overline{\lambda}(i|h)\}, \{\overline{\beta}(w|i,\overline{h}_i)\})$$

$$(7.65)$$

$$= \sum_{h,w} N(h,w) \sum_{i} \frac{\lambda(i|h) \cdot \beta(w|i,\overline{h}_{i})}{\sum\limits_{i'} \lambda(i'|h) \cdot \beta(w|i',\overline{h}_{i'})} \log \left[\overline{\lambda}(i|h) \cdot \overline{\beta}(w|i,\overline{h}_{i})\right]$$

Taking the partial derivatives with respect to $\overline{\lambda}(i|h)$ and observing the normalisation constraint, we obtain the equation:

$$\overline{\lambda}(i|h) = \frac{1}{N(h,\cdot)} \sum_{w} N(h,w) \frac{\lambda(i|h) \cdot \beta(w|i,\overline{h}_{i})}{\sum_{i'} \lambda(i'|h) \cdot \beta(w|i',\overline{h}_{i'})}$$
(7.66)

In this form, the interpolation parameters $\overline{\lambda}(i|h)$ depend on the history h the number of which can be large. Therefore, some sort of tying might often be useful (Jelinek and Mercer 1980).

Taking the partial derivatives with respect to $\overline{\beta}(w|i, \overline{h}_i)$ and observing the normalisation constraint, we obtain the equation:

$$\overline{\beta}(w|i,\overline{h}_i) = \frac{\sum\limits_{h\in\overline{h}_i} N(h,w) \frac{\lambda(i|h) \cdot \beta(w|i,h_i)}{\sum_{i'} \lambda(i'|h) \cdot \beta(w|i',\overline{h}_{i'})}}{\sum\limits_{w'} \sum\limits_{h\in\overline{h}_i} N(h,w') \frac{\lambda(i|h) \cdot \beta(w|i,\overline{h}_i)}{\sum_{i'} \lambda(i'|h) \cdot \beta(w|i',\overline{h}_{i'})}}$$
(7.67)

7.8.3 Absolute discounting and backing-off

The basic idea is to subtract a constant value from the counts of the observed events. Thus we define the model for absolute discounting and backing-off:

$$p(w|h) = \begin{cases} \frac{N(h,w) - b_h}{N(h,\cdot)} & \text{if } N(h,w) > 0\\ b_h \cdot \frac{W - N_0(h,\cdot)}{N(h,\cdot)} \cdot \frac{\beta(w|\overline{h})}{\sum\limits_{w':N(h,w')=0} \beta(w'|\overline{h})} & \text{if } N(h,w) = 0 \end{cases}$$
(7.68)

Using the same manipulations as for linear discounting, i.e. separating the singletons, ordering and carrying out the sums, we have for the leaving-oneout log-likelihood function:

$$F(\{b_h\},\{\beta(w|\overline{h})\}) = \sum_{h} \left[N_1(h,\cdot) \log b_h + \sum_{r>1} r N_r(h,\cdot) \log [r-1-b_h] \right] + \sum_{w} N_1(\cdot,w) \log \frac{\beta(w|\overline{h})}{\sum_{w'} \beta(w'|\overline{h})} + const(\{b_h\},\{\beta(w|\overline{h})\})$$
(7.69)

For the backing-off distribution $\beta(w|\overline{h})$, we obtain the same equation as for linear discounting. For the discounting parameter b_h , we obtain the following equation after separating the term with r = 2:

$$\frac{N_1(h,\cdot)}{b_h} - \frac{2N_2(h,\cdot)}{1-b_h} = \sum_{r\geq 3} \frac{rN_r(h,\cdot)}{r-1-b_h}$$
(7.70)

For this equation, there is no closed-form solution. We will derive an upper and a lower bound for b_h . Obviously, the right hand side is always greater than or equal to zero, and thus we get the upper bound:

$$b_h \le \frac{N_1(h, \cdot)}{N_1(h, \cdot) + 2N_2(h, \cdot)}$$
(7.71)

Lower bounds are obtained by observing the inequalities

$$\sum_{r\geq 3} \frac{rN_r(h,\cdot)}{r-1-b_h} \leq \sum_{r\geq 3} \frac{rN_r(h,\cdot)}{r-2}$$

Using the additional inequality $b_h(1-b_h) \leq 1/4$; we have the lower bound:

$$\frac{N_1(h,\cdot) - \frac{1}{4} \sum_{r \ge 3} \frac{rN_r(h,\cdot)}{r-2}}{N_1(h,\cdot) + 2N_2(h,\cdot)} \le b_h$$
(7.72)

Using the inequality

$$\sum_{r \ge 3} \frac{rN_r(h, \cdot)}{r-2} \le \sum_{r \ge 3} rN_r(h, \cdot) = N(h, \cdot) - N_1(h, \cdot) - 2N_2(h, \cdot)$$

this can be simplified to the more convenient bound:

$$\frac{N_1(h,\cdot) - \frac{1}{4} [N(h,\cdot) - N_1(h,\cdot) - 2N_2(h,\cdot)]}{N_1(h,\cdot) + 2N_2(h,\cdot)} \leq b_h$$
(7.73)

8 Physical characterisation and description

8.1 Introduction

Why should a handbook about standards and resources for spoken language systems bother its readers with electroacoustics or other physical or technical basics? We consider this to be worthwhile for several reasons. Firstly, working on speech data will increase reliability of results on the basis of awareness of the basic physical facts. This holds true especially for recording speech data. Furthermore, by agreeing on common standards results become comparable. On the other hand, speech data and results also become more valuable to the speech community. Finally, by sticking to standards or recommendations, those with little technical training may also achieve their goals by finding efficient ways of selecting and using appropriate tools. Therefore the aim of this chapter is to motivate the reader to concern himself with the physical background. In a single chapter it is impossible to cover the large range of possible speech applications or to deal with all important physical aspects that may arise in this context, and further explanations and assistance may be found in the literature recommended at the end of each section.

The remainder of this section introduces the concept of the communication chain from the production of speech to its exploitation. Recording specifications should be as goal-directed as possible, and therefore we distinguish between an "ideal" (task-independent) and a "real-life" (task-dependent) approach. Additionally, specific requirements are mentioned in the contexts of building speech corpora (Chapter 3) and speech assessment (Chapter 10). Section 8.2 presents the part of the basic terminology and common notations that a person working on speech data is likely to be confronted with. In Section 8.3 the human elements of the communication chain are investigated; here we start with the question of how a talker or listener can be characterised and how he should be selected for a specific task. Then the requirements on the minimum recording chain are presented in Section 8.4: what kind of microphone should be chosen, what are the influences of the recording environment and how we get the speech data, including optional parallel recordings capturing fundamental frequency, physiological data or mimics and gestures. Being more speaker-oriented, Section 8.5 deals with the conditioning of a speaker in a natural or artificial auditory and visual environment. In Section 8.6 and 8.7 the technical elements of the recording chain are investigated once more, focussing on linear or non-linear distortions and on reproducibility assurance procedures, respectively. The final

section presents some tools for further task-specific processing of speech data, aiming at signal analysis, measurement, and conditioning.

8.1.1 The communication chain

Due to the different possibilities of producing, processing and exploiting spoken utterances there exists a tremendous variety in the architecture of the so-called communication chain. Therefore we define the *communication chain* as the connection(s) between a talker and a listener via an auditory, a visual and/or an electric channel. While these are parallel channels of information flow, the electric channel as well might be seen as serial and/or parallel connections of electric devices and channels. Figure 8.1 may give a rough impression of this somewhat simplified scheme.



Figure 8.1: Scheme of the communication chain

This scheme consists of the following elements:

- 1. We have to recognise the talker (source) and the recording environment as a whole.
- 2. The auditory and visual environmental factors both have impact on the talker's behaviour as well as on the probes at the sensor's position.
- 3. We have to be aware of this talker-environment feedback, although the communication chain suggests a unidirectional talker-to-listener succession.
- 4. To collect the talker's data of interest (i.e. speech, lip movements or glottal frequency), sensors and transducers are to be applied: the latter converts

acoustic and mechanical energy into an electrical signal (like a microphone) or vice versa (like loudspeakers or headphones).

- 5. If the electrical signal has to be analysed or manipulated, this can be done by signal processing devices like filters, amplifiers, A/D-converters, computers with room simulation algorithms, et cetera.
- 6. The processed signal may be played back via loudspeakers, sent to a storage device or transmitted via an information channel like the telephone line.
- 7. The listener makes up the end (sink) of the communication chain.

In this chapter we differentiate between two opposite strategies for the actual specification of a communication chain. The first strategy, called the ideal or flawless approach, tries to capture the speech signal as cleanly as possible in a domain and scenario independent way. The advantage is that these data may be applied to many tasks with "average" suitability, without being ideally adapted to their specific conditions. Another advantage is given by flexibility in exploiting the same data: many post-processing possibilities exist so that many task-specific signal characteristics may be imposed after the recording itself. But the talker's conditions and some environmental factors are also reflected by the "clean" data, and the possibilities of subsequent corrections or manipulations are limited. To yield so-called *flawless speech*, we have to consider the dilemma of motivating a natural way of speaking on the one hand and optimising the more technical circumstances of the recording session on the other hand. One has to come to a minimum set of decisions: what kind of speaker (cf. Section 8.3), what kind of auditory and visual environment (cf. Section 8.5) and how to capture the speech signal in an optimal way (cf. Section 8.4).

The opposite recording strategy may be called a real-life or on-site approach: From the beginning the communication chain is adapted to a specific scenario as closely as possible. For instance, if a speech recognition device that makes up a part of an information system for in-car inquiries via a mobile phone is to be evaluated, the speaker has to sit in a moving car, he has to drive the car himself and the speech data have to be transmitted over the wireless telephone network. As in similar cases, the simulation of the acoustic environment is not the crucial point, but the situation-dependent speaking style influences the resulting speech signal significantly. Within this approach we find the dilemma of ensuring real-life conditions on the one hand while performing the recording in an optimal way on the other hand.

8.1.2 Specifications for speech corpora production and speech assessment environments

Both approaches are suitable for the generation of spoken language corpora (cf. Chapters 3 to 5): while flawless speech may be sufficient for allround utilisation, a collection of in-car speech at certain conditions (car speed,

road surface, etc.) may be highly valuable for the example given above, but is restricted to this special case. In both the ideal and the real-life approaches, a detailed protocol of all recording conditions is mandatory (cf. Section 8.7 on reproducibility).

In designing speech assessment environments, we can find a similar distinction between field and laboratory tests (cf. Chapter 12). While a field test is executed in a real (natural or virtual) environment, laboratory tests with subjects may utilise both clean and real-life speech. For speech output assessment in case of speech synthesisers or codecs, the "clean" output speech has to be subsequently adopted to real-life conditions by signal processing steps (cf. Section 8.8).

8.1.3 Recommendations on recording strategies

Let us summarise the introduction by the following recommendations:

- 1. Try to follow one of two strategies: consider either all possible task-specific recording conditions ("real-life speech") or as few deviations from "standard" high-quality conditions as possible in order to gather speech signals that are as clean as possible ("flawless speech"). Both strategies can prove to be "optimal" for specific tasks, but a compromise will not.
- 2. In case of flawless speech: artificial recording conditions should not affect the speaker's behaviour.
- 3. In case of real-life speech: keep a record of all recording conditions, including the task-specific ones.
- 4. Consider the relevant recording and signal processing standards and recommendations.

8.2 Basic notations and terminology

8.2.1 Periodic and aperiodic motion

The simplest kind of motion is the sine wave, which is approximately the natural motion of a weight that bobs up and down on a spring swinging at a moderate displacement. In case of an undamped motion, the weight repeats the same activity with every *period* T. A related term is *frequency*, $\frac{1}{T}$, the number of periods in a given interval of time. For instance, a sine wave with a period of $\frac{1}{1000}$ second has a frequency of 1000 Hz or 1 kHz (with "Hertz" or "Hz" for cycles per second, "kiloHertz" or "kHz" for 1000 Hertz). Other characteristics of a sine wave are its *amplitude*, which determines the displacement from a reference point, and its *phase*, which refers to the relative displacement in time between sine waves of the same frequency. The amplitude as a function of time can be a measure of distance, as in the case of the spring, or of current or voltage in case of an electrical sine wave. The amplitude of a sound wave like speech is measured in sound pressure fluctuations above and below normal atmospheric pressure.

While pure sine waves are very rare in the real world of sound, they are the basic elements for more complex sounds like a bowed string or a sustained vowel: any repetitive (periodic) waveform can be expressed by an ensemble of sine waves, beginning with a fundamental wave and adding a set of harmonically related sine waves, whose periods are related as $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, and so on (*Fourier's Theorem*). Therefore every complex periodic time wave can be represented by the relative strengths of its fundamental wave and its harmonics, called a *frequency spectrum*. Analysis of a complex periodic time wave into its spectral components is known as *Fourier analysis*, and the procedure involved is the (*Discrete*) Fourier Transformation (DFT); specially optimised versions of the DFT algorithm are known as *Fast Fourier Transformation* (FFT) algorithms.

Another kind of sound is of aperiodic nature, like frication noise, and therefore has neither a period nor a fundamental frequency. Just as repetitive waveforms can be made up of harmonically related sine waves, noise can be represented by a continuous band of an unbounded number of sine waves. Sound containing all frequency components up to a limiting frequency with equal energy is called *white noise*.

8.2.2 The dB scale

The frequency range of audible sound is nominally 20 Hz to 20 kHz (a ratio of 1 : 1000), and people with normal hearing can perceive an amplitude range of sound pressure between the threshold of hearing $(20\mu Pa, \text{ i.e. micropascal})$ and the threshold of perception of 1 : 1000000. To reduce this large range of physical values to a smaller range of numbers the *decibel* (*dB*) scale was introduced.

The dB is not an absolute measuring unit. Rather it is the ratio of a measured value to a fixed reference value. The sound pressure level (SPL) in dB, indicated as dB-SPL, uses a logarithmic scale with the threshold of hearing defined as 0 db:

power ratio
$$(dB) = 20 \log \frac{P}{P_0}$$
 (8.1)

where $P_0 = 20\mu Pa$. With every multiplication of the sound pressure P by the factor 10, we add 20 dB to the sound pressure level: $200\mu Pa$ corresponds to 20 dB, $2000\mu Pa$ corresponds to 40 dB etc. By using decibels, the range of 1 : 1000000 is mapped onto a range from 0 dB to 120 dB: 120 dB-SPL may be emitted by aeroplanes during take-off, and the range of speech is between 30 and 70 dB, with normal conversation (1m separation) at 62 dB. In the case of two sound sources, the logarithmic scale means that the power levels in dB cannot be simply added to get the resulting level. While, for instance, 90 + 96 and 97 + 97 add to 97 and 100 db, respectively, the sum approximates to the larger value alone if the two levels differ by more than 10 dB.

8.2.3 Phones and sones

What do we hear? The logarithmic sound pressure level in dB-SPL does not correspond to the subjective level of perception. However, the latter is called loudness (measured in *sones*) and is roughly proportional to the 0.6th power of the sound pressure, within a wide amplitude and frequency range. The definition of loudness is based on the loudness level (measured in *phones*). While a sound intensity ratio can only be described by the loudness measure, the determination of loudness levels in a physical space leads to the so-called *equal loudness contours* or *isophones* for different intensity levels as a function of frequency (cf. *isobars* as pressure contours). Although the averaged isophones are standardised, originally both measures have to be determined by listening tests with subjects using a standard reference stimulus. For this stimulus (1000 Hz sine wave) the loudness level in phones is defined to be equal as the sound pressure level in dB.

In case of speech signals, the rule of thumb is as follows: doubling the sound pressure increases the sound pressure level, and also the loudness level (in phones) at 1000 Hz, by 6 dB. On the other hand, doubling the loudness (in sones) – and thereby the degree of perception – is equivalent to an increase of the sound pressure level by 10 dB.

The loudness level can be estimated on the basis of the frequency spectrum, taking into account the frequency response and the masking properties of the auditory system. Frequency response means that the sensitivity of the auditory system varies with frequency: the human ear is most sensitive at frequencies between 2000 and 5000 Hz, and most insensitive at low and high frequencies. This phenomenon is more distinct at low than at high sound pressure levels. Additionally, masking properties have to be considered, which means that the estimation of these measures for real-life sounds like speech is by far more complex due to time and frequency-dependent suppression or masking of signal components by the human ear.

8.2.4 Analog and digital speech signal representation

Today most signal processing and transmission techniques are carried out in the *digital signal domain*, with the advantages of greater robustness compared to the analog domain and of the flexible (programmable) and reliable use of digital computers and associated digital hardware to arbitrary accuracy. Therefore recommended that digital data should be used wherever possible. In many cases it is necessary to switch into the digital domain, for instance for storing speech data on Digital Audio Tapes (DAT's) or CD-ROMs, or for further processing on computers. In order to understand the characteristics and limits of digital signal representation, the basic concepts of sampling and quantisation must be understood.

8.2.4.1 Sampling

An analog signal is continuous in both time and amplitude. The transition to a time-discrete but amplitude-continuous signal is performed by the sampling process: by taking one amplitude value (or sample) every T_s seconds the original waveform is converted into a train of pulses. This signal representation is called *Pulse Amplitude Modulation (PAM)*, and all coding methods that try to reconstruct this pulse train are called waveform coding. T_s stands for the sampling interval, and $f_s = \frac{1}{T_s}$ stands for the sampling rate or sampling frequency. For choosing T without any loss of information, the sampling theorem has to be borne in mind: a band-limited analog signal may be represented by time-discrete sampling values at constant time intervals $T_s = \frac{1}{f_s}$ without any information loss if $f_s \ge 2f_{max}$, with sampling rate f_s . This is only defined for low-pass signals below a specified cut-off frequency, with spectrum $S_{j\omega} \equiv 0$ for $|\omega| \geq \omega_{max} = 2\pi f_{max}$) This means that the highest frequency component f_{max} in the analog signal to be sampled has to be lower than half of the sampling rate f_s . If you are unsure, this has to be guaranteed by low-pass filtering of the signal before starting the sampling process. Otherwise, the analog signal cannot be reconstructed from the samples without severe errors commonly called aliasing.

While PAM offers time-multiplexing capabilities, the pulse amplitudes are still sensitive to noise.

8.2.4.2 Quantisation and coding

In a second step, quantisation, sample amplitudes are binary coded with a binary word length of w bits per sample, in order to achieve an amplitudediscrete representation (linear Pulse Code Modulation or Lin-PCM). Consequently, the most similar value of 2^w possible amplitude values has to be chosen, the difference compared to the original amplitude being the quantisation error. While the bandwidth requirements increase by coding w bits (pulses) per sample, the digital signal is resistent to added noise distortions if the noise does not exceed one quantisation step and if the signal amplitude does not exceed the maximum discrete amplitude range. In order to waste no quantisation steps or bits, the recording level has to be controlled to take advantage, without overload, of the full recording range. The quality of linear PCM is commonly described by the signal-to-noise ratio SNR, referring to signal power and noise power.

In addition to this linear time-invariant coding of the original sampled signal, various modifications have been proposed to take full advantage of the long-time or short-time characteristics of the speech signal (Rabiner and Schafer 1978). One method of these is *logarithmic PCM* (A-law or μ -law Log-PCM), which uses a higher quantiser resolution at small signal amplitudes and larger quantisation steps at high amplitudes. Another improvement can be achieved by permanently adapting the range of the quantiser to the short-time signal amplitude. A different category of so-called *parametric coding strategies* applies assumptions about the speech production process within an "intelligent" speech coder, thereby shifting the costs from the transmission line (where very low bit rates can be achieved) to the signal analysis and synthesis stage.

For further reading consult O'Shaughnessy (1987); Rabiner and Schafer (1978); Pierce (1991).

8.3 Talker / listener descriptors

What is the ideal test population like? It should be representative (cf. Chapters 4 and 9).

There are two ways of collecting representative test populations:

- 1. to pick a large number of subjects (at least 1000) to have some statistically significant coverage;
- 2. to generate meaningful descriptors and to select an equally distributed population.

The advantage of the latter strategy is that the number of subjects to be tested can be reduced drastically in comparison with the purely statistical approach. Furthermore, it is always possible to select a specific subset from the whole population in order to focus the investigations on specific problems. This has to be traded off by some extra investment in the clarification of the set of meaningful descriptors as well as by extra effort to be spent on the classification of each subject according to these descriptors.

For the statistical approach, however, there is no need for anything like an explicit classification; this may be crucial for on-site clandestine speech recordings.

Independently of the purpose for which a population is selected, listeners or talkers, its descriptors may be divided into two types: general descriptors and task specific descriptors.

For the reasoning concerning which descriptor to select and why, consult Section 3.5. Here we focus on how to extract the relevant physical characteristics.

It is important to note that most of the physical descriptors, e.g. voice quality or precision of articulation, are long-term functions of time. Consequently, test populations are, generally speaking, subject to aging.

8.3.1 General (demographic) descriptors

Since the individual talking and listening properties of a subject strongly correlate with his social and demographic status, it is generally advisable to collect information on all the available relevant descriptors. These might comprise age, sex, education, profession, residence, nationality, ethnic affiliation, religion, etc.
8.3.2 Task specific descriptors

Task specific descriptors we define to be such that they directly describe the ability of a subject to perform a specific task. For example in listening tests it is crucial to check whether the experimentee is able to hear at all.

8.3.2.1 Talker descriptors

8.3.2.1.1 General talker descriptors

1. Voice / speech related medical records

Since certain diseases like inflammation of the vocal cords are known to potentially harm or at least influence the voice permanently, we recommend asking for any related medical records.

2. Voice / speech relevant habits

Voice relevant habits we consider to be smoking, drinking (cf. Section 3.5.2.6), and whether and to what extent a subject has received voice training or is accustomed to professional public speaking. Moreover, one should check whether a talker practises regular singing for either private or professional reasons, and how much.

8.3.2.1.2 Anatomical (voice) descriptors

With anatomical talker descriptors, we distinguish between descriptors derived from the subject's laryngeal behaviour and miscellaneous descriptors. These types of descriptor are not mutually exclusive, i.e. a laryngeal feature might well be explained in terms of jitter, shimmer, or glottal-to-noise excitation and vice versa.

The difference is that a close look at the laryngeal properties of voice necessitates the use of special pitch determination instruments (cf. Section 8.4.2.1), whereas the other descriptors rely on the analysis of the microphone time signal.

A perceptual classification based on listening might be sufficient, as long as the classification is performed consistently by the same judge(s) on the entire population.

But there exists no such thing as an absolute and generally accepted scale for the quantisation of voice quality.

1. Laryngeal descriptors

The following explanations on how to qualify a voice are partly based on a signal output by a so-called laryngograph (cf. Section 8.4.2.1). This signal, which is proportional to the electrical impedance of the larynx (i.e. the opening/closure of the glottis), is referred to as Lx; Fx denotes the fundamental frequency (a direct derivative of Lx), and Cx stands for the scatterplot of $Fx' = Fx + \delta Fx$ over the fundamental frequency (Fx). The latter represents a measure of the variance of the fundamental frequency as a function of the fundamental frequency.

BREATHY VOICE: A breathy voice results from slow, sometimes incomplete closure of the vocal folds during the laryngeal cycle. It is more often found in women than in men. The auditory impression is that of a "gentle" voice, which in women sometimes reaches the point of sounding "whispery".

A more sinusoidal shape of the Lx signal, as well as a lower closed/open phase ratio calculated from it, are an indicator of breathy voice. Acoustically, the zero-crossing rate in the 3–4 kHz band during the voiced sections of an utterance is a measure of accompanying glottal friction. A further measure is the relative strength of the first and second harmonics (there being a step-down from the first to the second rather than an equal slope).

HARSH VOICE: This is the converse of a breathy voice, and is more often found in men. It results, probably, from a very fast closing gesture, and a high closed/open phase ratio. It is the sort of voice that "carries" well in voice babble.

As might be expected by its converse relation to breathy voice, the more vertical closing phase of the Lx wave and the higher close/open phase ratio indicates a harsher voice. Acoustically, the absence of the step-down from the first to the second harmonic and an overall flatter spectrum are characteristic of the voice quality.

CREAKY VOICE: This is the result of irregular laryngeal vibrations, often with a cycle of "normal" duration being followed by a cycle of roughly twice the normal duration. It is found in both men and women. In some speakers it occurs at particular parts of an intonation contour, typically at the end of a phrase, when the voice sinks to the bottom of its range.

Irregular laryngeal vibrations are clearly visible in the Lx, and the Fx distribution reveals a clear secondary mode about one octave below the main mode. There are usually also points on the Cx scatterplot to either side of the main diagonal at the lower end of the speaker's frequency range.

HOARSE VOICE: This adjective is often given to a mixture of laryngeal irregularity with breathiness. In everyday terms it is the sort of voice that makes you think the speaker has been shouting a lot.

The combination of laryngeal irregularity and glottal friction in this voice quality means that it is open to both laryngographic and acoustic representation. The points to the side of the main diagonal of the Cx plot usually spread along the whole of the speaker's range. Acoustically, a similar zero-crossing measure can be used as for breathy voice. In addition, however, the irregularity is clearly visible on a spectrogram.

2. Miscellaneous descriptors

It is currently impossible to give an all-embracing compendium on voice descriptors. To this end, we restrict ourselves to those we consider to be most common:

- 298 Part II: Spoken language characterisation
- VOCAL TRACT SIZE: It is generally agreed that body size correlates with vocal tract size. However, observation of head size relative to body size is a further criterion. We recommend logging such personal data, i.e. height and weight and head perimeter of all subjects.

First and third-formant averages over a given utterance, spoken by persons with the same regional accent, can be used as an indicator of relative vocaltract length.

JITTER AND SHIMMER: Jitter and shimmer are measures of the average perturbation of someone's fundamental frequency and of its magnitude, respectively. They are given by the formula:

$$PF = \frac{1}{N-1} \sum_{n=1}^{N-1} \left| \frac{u(n) - u(n-1)}{u(n)} \right|$$
(8.2)

where u(n) denotes either the length of the observed excitation period (jitter) or the energy in the period (shimmer). Details on how to extract a value for u(n) may be found in Kasuya et al. (1993) and Michaelis and Strube (1995).

Both measures mutually correlate to a high degree and have to be expected to have high values in creaky as well as in hoarse voices (see above).

GLOTTAL-TO-NOISE EXCITATION PARAMETER: The glottal-to-noise excitation parameter (GNE parameter) gives a figure of whether vocal excitation is mainly due to glottal vibration (GNE = 1) or rather turbulent noise (GNE = 0).

Since it is a measure of harshness it exhibits high values in harsh voices and it will be found to be small in breathy and hoarse voices. For further details on this parameter consult Michaelis and Strube (1995).

8.3.2.1.3 Habitual speech descriptors

1. Average level and dynamics of rate of articulation

This can be quantified by average word length for a number of agreed isolated words. In continuous speech, average duration of (underlying) syllables in a given utterance, excluding pauses, may serve. This allows for a rate measure which excludes consideration of articulatory precision. At the same time the minimum and maximum duration of the syllables can be recorded to establish a measure of the dynamics of the rate of articulation.

2. Precision of articulation (coarticulation)

It is difficult to define this in objective terms, and possibly there will be disagreement in selecting speakers, except for extreme cases. Note that this is not necessarily the same as speaking slow or fast, though the two dimensions may covary among the same speakers. Though it has not been investigated experimentally, we can assume in the first instance that the impression of precise articulation has to do with the consistent avoidance of frication for stops, and not producing fricatives as approximants, not eliding or slurring unstressed syllables very much. These are undoubtedly properties that are of interest with respect to recogniser assessment. In contrast to the rate of articulation, this measure should be based on the average duration of actually realised syllables. Elided syllables, which contribute to the rate measure, would therefore be ignored.

3. Average level and range (dynamics) of fundamental frequency

Though the fundamental frequency in principle is a function of the subject's anatomical data, it is modulated in both directions by intonation, tone and accentuation. The dynamics are constituted by the maximum and the minimum frequency observed in an agreed set of utterances.

4. Average level and dynamics of speech intensity

The intensity contour provides speech with what is commonly known as volume and rhythm, respectively. It can be derived directly from the energy of the time-signal of the recording. As with other measures of this kind, this should be done on an agreed set of utterances.

8.3.2.1.4 Audiometric descriptors

The manner of speaking and the very ability to speak depends on the ability to hear. For that reason it is recommended to check for potential hearing impairment of subjects to be recorded, at least in case of doubts. Appropriate tests are given by pure-tone audiometry and so-called speech audiometry (cf. Section 8.3.2.2.2). Audiometric "functionality" of talkers becomes crucial in recordings in which acoustic feedback or stimulation is planned during the recording.

8.3.2.2 Listener descriptors

8.3.2.2.1 General listener descriptors

1. Hearing related medical records

Various diseases, such as inflammation of the middle-ear can significantly degrade hearing properties, even if they occurred decades earlier. For this reason we recommend asking potential candidates if they happen to suffer from any such disease, and to ask for the anamnesis.

2. Hearing relevant habits

- AVERAGE NOISE CONSUMPTION: The kind and amount of noise a subject is frequently exposed to gives a clue to possible hearing losses as well as to the degree to which he is accustomed to noisy environments. A person, for example, who is used to professionally communicating in noisy environments exhibits significantly better listening performance than inexperienced listeners. The average noise is measured by its level, duration and spectral characteristics. A comprehensive discussion on the judgement of effects and figures of everyday noise loads can be found in Rose (1971).
- EXPERIENCE: Experience in listening experiments clearly enhances performance in such tests. For this reason we recommend always establishing a "listening test record" for all members of a test population. Primarily this should include the types of test the subject is experienced in.

A "normalisation" of all listeners with respect to experience, however, can generally be achieved by giving dummy examples prior to the actual experiment.

8.3.2.2.2 Audiometric descriptors

1. Pure-tone audiometry

Pure tone audiometry provides a measure of the hearing sensitivity as a function of frequency. It is measured by air conduction and by bone conduction. In addition to the absolute sensitivity in dB-SPL, a pure-tone audiogram also displays how much a subject deviates from the average listener and whether this deviation is within an admissible range or not.

The technical setup and procedure is standardised to a high degree (ISO 1964) and appropriate test equipment is widely available.

For further details on pure-tone audiometry consult Rose (1971).

2. Speech audiometry

The goal of speech audiometry is to investigate the listener's response to speech. Despite theoretical and practical difficulties, it provides a method by which such assessment can be made. It is almost too obvious to state that it is not normally necessary to attend to or discriminate among puretone stimuli, but rather it is constantly necessary to identify speech units. Speech audiometry is concerned with answering the three questions:

- What is the lowest intensity level at which a listener can identify simple speech fragments?
- How well does a listener understand everyday speech under everyday conditions?
- What is the highest intensity at which the listener can tolerate speech.

Though not really standardised, certain standard procedures for the speech audiometry have been established during recent decades. For comparison between separately tested populations, however, the exact test configuration and procedure has to be recorded.

For further reading consult Barry and Fourcin (1990); Kasuya et al. (1993); Michaelis and Strube (1995); Rose (1971).

8.4 Transducer characteristics

The task of recording speech requires the operator to make a chain of decisions starting with the appropriate microphone and ending with the choice on what coding might be the right one for the specific purpose. Even prior to that, a decision has to be made on what recording environment will be suitable, i.e. whether studio recordings are required in order to obtain speech which is as "flawless" or "clean" as possible, or on-site recordings, which provide a rather natural talking situation for the subject.

The major concern in both recording modes (studio/on-site) is to avoid further degradation of the quality of the speech once one has obtained technical control over the signal – either by the remaining recording chain or by the way the data is sampled, coded or stored. The difference between both is constituted by the fact that in on-site situations, such as in telephone recordings, control is very limited, whereas in studio recordings everything from the microphone to the storage device can be determined in advance. Consequently, this section concentrates on giving recommendations on the minimum requirements a recording environment should fulfil for the recording of technically "flawless" or "dry" speech. *Flawless speech* we define as

the unweighted reproducible 1:1 transduction of an acoustic signal emitted by a speaker into a sequence of 2 byte numbers that is free of any room or environment information, exhibits a sufficient signal-to-noise ratio of at least 50 dB, and has been produced under recording conditions that do not impose any stress upon the speaker in addition to what might be intended for a given talking situation (see Chapter 4).

In so far as the recording manager is able to exercise control over any recording component, everything said about studio recordings also directly applies to on-site recording situations.

Further, we describe recording techniques that might be employed parallel to the pure speech recording, such as pitch determination by laryngograph, physiological measurements, or mimicry and gesture recording.

8.4.1 Microphones and speech recordings

8.4.1.1 Microphones

The requirements concerned with the choice of the right microphone for a given application can be summarised as follows:

- electroacoustic performance, such as sensitivity, frequency range, transient response, non-linearities;
- mechanical characteristics, such as size, robustness, etc.;
- electrical characteristics in view of compatibility with other equipment;
- insensitivity to external influences such as shock, vibration, electric and magnetic fields, wind;
- cost, handling, and other external aspects.

For speech recording purposes under laboratory conditions, the requirements to be focussed on are the flattest possible frequency response and a specified type of directivity to be as constant as possible over the intended frequency range.

8.4.1.1.1 Conversion principles

Basically, there are two different physical effects most microphones use to convert acoustic energy into electric energy. Consequently, there are two major groups most microphones can be categorised into, depending on their functional principle.

1. Dynamic microphones

Dynamic microphones use a constant magnetic field to induce voltage in a moving coil mechanically coupled to the diaphragm. Since the output voltage of the microphone is directly generated by the conversion process, no external power supplies are required. Dynamic microphones are quite robust and may be exposed even to high sound pressure levels, which makes them suited for close-talking applications, for example in headsets. The major disadvantage of the dynamic operation principle is that in addition to the diaphragm the comparably heavy moving coil also has to be moved by the sound pressure, resulting in a poorer transient response of the microphone. For this reason dynamic microphones are, with some exceptions, rarely used as top quality studio microphones.

2. Condenser microphones

Condenser microphones basically consist of a capacitor, one of the electrodes of which is formed by a conductive membrane. This membrane is exposed to the incident sound and, when moved back and forth by the sound pressure, slightly changes the capacitance of the capacitor. When the load on the capacitor is kept constant the capacitance changes will, for the voltage across the electrodes, follow the movements of the membrane as long as the voltage changes are small compared to the total voltage across the electrodes. Since the membrane can be manufactured from very thin plastic film material with a conductive layer of vaporised gold or aluminium, it will follow the sound pressure quite exactly and the signal produced by the microphone will be a rather precise reproduction of the original course of the sound pressure. For high-quality studio recordings most microphones used are condenser microphones.

Since the output impedance of the condenser microphone is high, all condenser microphones contain an impedance converter to render an output impedance of approximately 200 Ohm. Therefore condenser type microphones need some kind of power supply not only for the impedance converter but also for the polarisation voltage across the electrodes prescribed by the operation principle. The usual way of supplying condenser microphones besides batteries is the use of a so-called phantom power supply which is connected to the output terminals of the microphone. The standard phantom power supply voltage is 48 V DC. To avoid DC offset on the speech signal, most studio microphones include an integrated optional highpass filter with passband beginning at a frequency slightly above 50Hz.

8.4.1.1.2 Directional characteristics

1. Omnidirectional microphones

The omnidirectional microphone is sensitive to sound without regard to the direction of the incidence. Thus it will pick up the wanted sound produced by the speaker as well as unwanted background noise. This feature makes an omnidirectional microphone a bad choice when unwanted noise sources are to be expected. On the other hand, it is the most simple type of microphone from the viewpoint of microphone design. As a matter of fact, omnidirectional microphones are the most natural microphones available since the least design compromises have to be made. Thus, omnidirectional microphones are the best choice for high-quality speech recordings as long as the ambient noise floor can be kept low. In addition, omnidirectional microphones do not exhibit the proximity effect. The proximity effect will be dealt with when considering unidirectional microphones.

2. Unidirectional microphones

The unidirectional type of microphone is most sensitive to sound arriving from one direction and more or less attenuates incident sound from other directions. Thus, unidirectional microphones will suppress intended sound when pointed at the wanted sound source, i.e. the speaker.

The construction of unidirectional microphones requires additional engineering effort if a flat frequency response is desired. This is due to the fact that unidirectional microphones respond to the pressure gradient of the sound field, which is frequency dependent. To compensate this dependence, additional tuning, either acoustic or electric, is required in order to yield a flat frequency response.

Moreover, unidirectional microphones show the so-called *proximity effect*. This effect occurs when spatially confined sound sources are to be picked up. The sound field of small sound sources may be approximated by spherical waves. The pressure gradient in a spherical wave is greater than the pressure gradient in a plane wave by a factor g:

$$g = \sqrt{1 + (\frac{c}{2nfr})^2}$$
(8.3)

where r denotes the distance between speaker and microphone, f is the frequency, and c the velocity of sound.

When r decreases, the second term in the equation increases and adds a frequency dependent component to the pressure gradient. Since the unidirectional microphone responds to the pressure gradient of the sound field, this behaviour yields a boosted bass response of the microphone at close talking distances which is termed *proximity effect*.

The proximity effect is generally unwanted except when recording musical instruments or vocalists, so that the increased bass response has to be compensated for by special microphone design with switchable bass-cut filters. In any case, the proximity puts constraints on the recording setup since it requires the speaker-microphone distance to be fixed when sound coloration is intolerable. The influence of the proximity effect decreases sufficiently when the talking distance is great enough, but this results in a decrease of sound pressure level which in turn has to be compensated for with additional gain at the microphone preamplifier, yielding a higher noise level.

There are several kinds of unidirectional microphone which are classified by the shape of their polar responses (Figure 8.2):

- 1. Cardioid microphones show best ambient noise suppression for incident sound from the back. Sensitivity loss is about 6 db at the sides of the microphone and 15-25 db at the rear.
- 2. Supercardioid microphones are least sensitive at 125 degrees off-axis, 8.7 db down at the sides and approximately 15 db down at the rear.
- 3. Hypercardioid microphones are least sensitive at 110 degrees off-axis, 12 db down at the sides and approximately 6 db down at the rear.

Typical applications for these types of microphones with respect to noise suppression are given below:



Figure 8.2: Typical polar patterns of various types of unidirectional microphone

- Cardioid microphones should be used when maximum attenuation is needed at the rear of the microphone.
- Supercardioid microphones should be used when a maximum difference between the front and the back-hemisphere is needed.
- Hypercardioid microphones should be used when maximum side rejection and the maximum rejection of reverberation and background noise is needed. Note that hypercardioid microphones show the greatest random energy efficiency, i.e. the greatest rejection of random-incidence sound.

3. Bidirectional microphones (figure-of-eight characteristics)

Bidirectional microphones are most sensitive at the front and at the rear. There is a plane of minimum sensitivity perpendicular to the direction of maximum sensitivity. This behaviour makes it most suited for the recording of more than one speaker. Bidirectional microphones should not be used to produce speech recordings from one speaker.

The bidirectional microphone also exhibits the proximity effect. The effect is approximately 6 db stronger as compared to cardioid microphones.

4. Ultradirectional microphones (shotgun)

The ultradirectional microphone is designed for distant pickup, e.g. in film or TV productions. It strongly attenuates off-axis sound by means of multipath interference at a long slotted tube mounted in front of a unidirectional microphone. Compared to omni- and unidirectional microphones the sound quality is relatively poor since it has been traded against good directivity. The ultradirectional microphone is not recommended for high-quality speech recordings.

5. Pressure zone microphones

A pressure zone microphone basically consists of an omnidirectional microphone mounted close to or into a boundary surface. The distance to the surface is significantly shorter than the wavelength given by the highest frequency to be picked up. Thus, the incident and the reflected sound will always interfere constructively, i.e. there are no comb filter distortions with this type of microphone. The directional characteristic of a PZM¹ is basically spheroid but the pickup range is limited by the boundary surface to a semisphere. The PZM microphone is recommended for recording situations in which the talker has to sit at a table.

$6. \ {\rm Headsets}$

The use of a headset microphone is recommended in all situations where a high ambient noise rejection is needed. The noise rejection properties are mainly due to the extremely close talking distance which allows preamplifier gain to be greatly reduced. Additional noise rejection can be achieved by choosing microphone capsules with directional properties. The good noise rejection behaviour has to be traded off by a degraded frequency response at low frequencies, which leads to an effect we already referred to as proximity effect (see page 304).

8.4.1.2 Recording environment

As already mentioned, a specific recording environment (see also Chapter 4) is either intended or not, depending on the underlying purpose the recording is to be made for. In the latter case, the environment itself as well as any physical feedback to the talker should be virtually non-existent with respect to the actual speech signal, i.e. acoustic feedback like noise, dialogues, or on-line instructions by the recording supervisor has to be conducted via headphones. It is necessary to control environmental conditions by avoiding any undesired room acoustics. Since then the talker has been deprived of his natural acoustic environment, with negative psycho-acoustic effects, some effort must be spent in making up for this (see Section 8.5.2).

For some purposes, e.g. basic phonetic research, when environmental impact on the talking subject is of no or little concern, the efforts can be limited to providing an appropriate "quiet" recording ambience (environment). For this, the number of other objects in the recording room apart from the talker himself (e.g. cameras, monitors, amplifiers, etc.), if they cannot be avoided at all, should be as small as possible. The objects should be kept as far away from the microphone as possible and, ideally, should be covered by acoustically absorbent material in order to keep unwanted and unreproducible reflections to a minimum. Furthermore, attention must be paid to the choice of the recording room itself.

 $^{^1\,\}mathrm{PZM}$ is a trademark of CROWN International Inc.

8.4.1.2.1 Small room acoustics

For the evaluation of recording spaces for high-quality speech recordings it is necessary to deal with some basic room acoustic properties. Since only few recordings are going to be made in large rooms such as concert halls, it is appropriate to deal with the acoustics of small rooms.

The distinction between large room acoustics and small room acoustics is necessary since it must be expected that the acoustic properties of a room vary substantially if its size becomes comparable to the wavelength (λ) of sound in the audible frequency range. The latter usually holds true for relatively small rooms such as those normally used for the production of speech recordings.

It is useful to analyse possible problems by looking at the eigenmodes (roughly, resonance properties) in rooms at different frequencies. Figure 8.3 shows that the frequency dependent behaviour of any room may be treated in four frequency ranges, where variable l denotes the longest dimension of the room and w is given as an empirical equation:

$$w = 1896\sqrt{\frac{RT}{V}} \tag{8.4}$$

with RT representing the reverberation time and V the volume of the room. At very low frequencies in region I the physical dimensions of the room are significantly smaller than the wavelength of sound. Thus, wave propagation is impossible in this frequency range and consequently the room acts as a pressure chamber in which the sound pressure does not depend on the probe position.



Figure 8.3: Closed room pressure zones

Region II is dominated by the first eigenmodes of the room, i.e. the wavelengths become comparable to the room dimensions. In this frequency region the acoustic properties of the room are best described by wave acoustics. Problems in this zone may arise due to constructive and destructive interference which will introduce comb filter effects when viewed in the frequency domain.

That is, when a sound source radiates sound in the frequency range given by region II, the sound pressure level that can be measured at different locations will extremely depend on the mode distribution in the room. At a fixed microphone position the measurable sound pressure level for a given frequency will depend on whether the standing waves will interfere constructively or destructively at that location.

Thus, in general, the acoustic transfer function between the sound source and the microphone position will not be flat but influenced by comb filter structures as depicted in Figure 8.4.



Figure 8.4: Typical comb filter structure

In large rooms, such as lecture or concert halls, frequency region II will lie well below the relevant frequency range for speech. This is not the case for rather small rooms, such as those often used for speech recordings. In such rooms, region II will often lie well within the speech frequency range, so that these rooms will need a large amount of well-designed acoustic treatment to be usable for the desired purpose.

In particular, the concept of reverberation time, known as a helpful measure from large room acoustics, will fail since the density of eigenmodes is not large enough and each mode has its own separable decay time.

Region III determines a kind of transition behaviour of the room and is dominated by diffraction and diffusion. The rules of wave acoustics have still to be considered, and when approaching the border to region IV, the rules of large room and ray acoustics begin to become valid. In region IV the wavelength of sound is substantially shorter than the room dimensions so that ray acoustics is a good tool for describing the behaviour of the room.

8.4.1.2.2 Recording rooms

1. Laboratory room

Speech recordings in typical laboratory environments are sometimes made in a kind of workbench situation when no special recording facility is available. Recordings made in a laboratory environment are often used to test speech recognition systems, as lab speech recordings seem to reflect best natural speech recognition situations, without requiring too much effort concerning the recording setup.

For standardisation purposes, however, the acoustic environment of a laboratory room is worst suited. Particularly when the recordings are made using a speaker sitting at a desk with the microphone being placed on the desk, the setup will lead to strong destructive interference due to reflections from the table surface.

In the frequency domain, this interference produces comb filter structures as shown in Figure 8.4, which will lead to periodic dips in the spectrum of the recorded speech signal. The frequencies where the dips can be found are dependent on the path difference of the direct and reflected sound and will strongly vary the sound coloration of the recorded speech signal when the speaker moves relative to the microphone or the table.

2. Soundproof booth

A sound-insulated and acoustically treated booth or small chamber is often used in clinical audiometry or in psycho-acoustic experiments. The advantage of this kind of equipment is that it is comparably inexpensive and may easily be standardised.

The kind of environment this equipment provides is, however, not recommended for high quality speech recordings for scientific purposes, since small rooms exhibit strong eigenmodes at relatively high frequencies which may lie well within the speech frequency region. Due to the small dimensions of the booth the acoustic treatment of the inner surface will generally not suffice to provide enough absorption for the resonances to disappear.

As a consequence, speech recordings produced in this environment will exhibit strong linear distortions, i.e. sound coloration.

3. Recording studio

Speech recordings may be made in a professional recording studio. The advantage of this type of recording environment is that it is widely available and that the recording location may be rented only for the recording

sessions. This will reduce cost, as the fact that the financial effort for the acoustic treatment of the studio will be restricted to the hiring fees. The major disadvantage of using a recording studio is that the recording conditions and especially the acoustic conditions are not standardised in any way. Moreover, it will generally not be possible to design the acoustic environment of the recording room according to the needs of speech recordings.

4. Anechoic chamber

The use of anechoic chambers for speech recordings is recommended from the acoustic point of view since it exhibits well defined acoustic properties. The almost total lack of wall reflections above a critical frequency, depending on the depth of the absorptive lining of the walls, renders the best approximation to free-field conditions in a noise-insulated environment.

The presence of free-field conditions is especially important with respect to the freedom of choice of the proper microphone to be used for recording. In most of the other recording environments discussed, the type of microphone to be used is largely influenced by the properties of the room, e.g. to suppress ambient noise or wall reflections and reverberation. For example, if a studio microphone with selectable directional properties is placed in an anechoic chamber, the sound of the recording does not depend on the selected directivity of the microphone.

Of comparable importance is the fact that the distance of the microphone relative to the speaker is least influential in an anechoic chamber since the microphone is always in the direct sound field of the speaker, and changing the distance only results in changes of the microphone output level as long as the proximity effect is negligible for pressure-gradient microphones.

Problems in the anechoic chamber may arise when a natural talker's response is to be elicited, e.g. in a dialogue situation, and when inexperienced speakers are used. These problems may arise due to the more or less unnatural perceptual effect which the anechoic chamber imposes on the subjects. For this reason, an appropriate form of acoustic feedback to the speaker that gives a natural room impression is highly recommended, especially for lengthy and psycho-acoustically sensitive recordings (see Section 8.5.2).

8.4.1.3 Recording chain: studio vs. on-site

8.4.1.3.1 Studio

The subject of this section is the minimum recording chain, i.e. the minimum number of mutually connected components that technically transduce the acoustic speech signal into a sequence of 16 bit numbers stored on digital memory media. As depicted in Figure 8.5, this basically comprises the

microphone itself, the preamplifier, the transmission line, and finally, the sampling device.



Figure 8.5: The minimum recording chain

For high-quality speech recordings, the overall noise figure, i.e. the signalto-noise ratio (SNR) of the setup has to be taken into special consideration. Assuming normal vocal effort and a talking distance of 30 cm, the SPL at the microphone capsule should rise to a level of about 75 db. If the recording takes place in an anechoic chamber, the ambient noise might level to about 20 dB-SPL such that the SNR at the front end of the recording chain equals to 55 db. All subsequent technical devices should be designed and connected to each other in such a way that this input SNR is degraded as little as possible. For a detailed discussion of noise figures and related terms and topics, please refer to Section 8.7.

1. Pop noise

Microphones used at close talking distances should be sufficiently protected against pop noise, generated for instance during the articulation of plosives. This is usually accomplished by either external or internal wind shields. Susceptibility to pop noises is strongly dependent on the microphone position.

The microphone should be removed from regions where considerable air flow is to be expected during articulation. A reasonable measure is to situate the microphone about 15 degrees off the direct talking axis.

2. Microphone

The internal noise of a microphone is usually given as the equivalent acoustic input noise level rendering the same output voltage. High-quality studio microphones show equivalent input noise (EIN) around 20 db-SPL, comparable to the ambient noise level in very quiet (sound-insulated) rooms. If added to the ambient noise of 20 db-SPL for such an anechoic chamber, this

reduces the output SNR to about 52 db. The equivalent noise level should be mentioned in the manufacturer's specification sheet.

3. Microphone preamplifier

The microphone preamplifier should meet standards for studio equipment. The minimal requirements for speech recordings are a flat frequency response, low distortion, low noise level, sufficient gain and a linear phase response.

The first two requirements are met by most preamplifiers built according to modern technology. The most serious problems occur in making a compromise between high gain and low noise. In general, noise generated within the preamplifier should not worsen the signal-to-noise ratio given by the EIN of the microphone. The input noise of high-quality microphone preamplifiers should be less than -125 dBu at 200 Ohm input impedance (dBu reference voltage: a0 dBu = $0.775 V_{eff}$ or 1 mW at 600 Ohm), which roughly corresponds to the thermal noise of a 200 Ohm resistor.

Usually, microphone preamplifiers allow the gain to be tuned from 0 to 60 db which is sufficient for microphone distances of about maximally 30 cm at a reasonably low noise level. Greater microphone distances, e.g. 50–60 cm, which may occur when the speech signal is picked up by a PZM microphone placed on a table in front of the talker, require amplifier gain in excess of 60 db which may result in audible noise during pauses.

4. Wiring and transmission lines

The requirement for absolutely correct wiring, i.e. sufficient shielding and grounding of the recording chain cannot be over-emphasised: this means connecting all components to a single solid ground. To avoid any unwanted induction into the transmission lines, these should possess a full-mantle shielding that has also been properly connected to the same ground (Figure 8.5). It is advisable to keep all lines as short as possible as well as to keep them away from any other electrical equipment.

It is standard in the high-quality speech-recording area to use balanced systems, i.e. to feed the speech signal into the recording chain along with its negative (180 degree phase shifted) counterpart (Figure 8.6).

Since both conductors in a balanced system pick up the same stray signal, noise that has eventually been induced to the system along the feedway can be cancelled out by the summation of the once more inverted signal with its unshifted double.

8.4.1.3.2 On-site

As stated previously, everything valid for the recording chain in a studio environment in principle holds also true for on-site recordings. A major



Figure 8.6: Noise cancellation on balanced microphone lines

difference, however, is that in real-life recordings, additional stages may be inserted into the recording chain which exhibit more or less unknown physical properties.

Telephone recordings

A very common method of speech data sampling is via the telephone. Hthat ere, only very general statements can be made on the quality of the microphone itself, not to mention the telephone network and the possibly intervening radio links when mobile phones are in use, or when communication takes place via satellite.

Recommendations on how to use a telephone can not be given. From the technical point of view, however, it has to be mentioned that the speech signal arriving at the receiving telephone has to be by-passed, sampled, and stored at some point prior to the acoustic output, i.e. it should never be captured by a microphone recording of the speech signal emitted from the telephone earpiece.

A coarse distinction between telephone networks can be made in terms of whether they use the analog or the digital signal domain. Whenever the operator has the choice, he should use digital telephone networks (in EU-Europe ISDN-network). This guarantees best possible signal quality in terms of noise and distortion. At the same time he must be aware of the fact that telephone networks may not be homogeneous in this respect, even within the same network.

Furthermore, the attention of operators must be directed to certain drawbacks of recording speech via the telephone:

• Frequency range is limited between 300 Hz and 3400 Hz in contrast to a natural speech frequency range of about 75 Hz to 8000 Hz. In consequence, this prevents proper pitch evaluation as well as sufficient spectral analysis

of high-frequency components such as those associated with fricatives, for example.

- In digital telephone networks speech dynamics are degraded to a limit of 42 db-SPL due to an 8 bit A-law coding (ISDN).
- Transmission properties change from network to network.
- Overseas communication is commonly transmitted via satellite. This potentially adds additional non-linear distortions (e.g. echoes) to the speech signal.

A discussion on whether telephone recordings are suitable for a specific purpose or not may be found in Chapters 3 and 4. Details of what kinds of distortion are imposed on a speech signal in a telephone network, and how to get a figure of their magnitude, are given in Section 8.6.

8.4.1.4 Data collection

For studio recordings, the data collection stage comprises the A/D-conversion of the analog audio signal and its storage on permanent memory media. We strongly recommend using digital data storage in general, and a hard disk directly connected to the sampling device (computer) in particular.

If all phonetically relevant information in a speech signal is spectrally restricted to a frequency range from 0 to $8.000 \,\text{Hz}$, the standard sampling frequency of A/D-converters for speech recording purposes, following the sampling theorem, is 16 kHz. Appropriate off-the-shelf equipment for speech sampling in real-time should be available for all current computer systems. These would include all filters necessary for proper preprocessing of the analog speech signal according to the sampling theorems. Attention has to be paid to the filters involved: these must be designed to be strictly linear in order to avoid unacceptable phase distortions.

The standard format of speech data is SHORT (16 bits, signed, linear) which corresponds to a representable value range of -32768 up to +32767, i.e. maximum recording dynamics of 96 db-SPL. With a properly calibrated microphone preamplifier at the front end, this should suffice for a peak factor in the recording session as well as the projected SNR of about 50 db at the microphone output.

Alternatively, a DAT (Digital Audio Tape) may be used to store the speech data. The standard sampling frequency is 48 kHz with a 16 bit resolution. This poses less strict requirements in view of the linearity of the filters involved. On the other hand it is rather cumbersome to access recordings made by a DAT for further processing.

When speech has been collected via a digital telephone network it might be necessary to resample the incoming signal according to the required sampling frequency of the recording station. On a digital recording device this is easily achieved by standard algorithms; if a DAT is used to record the digital signal, proper D/A-conversion is necessary. The easiest way to control the domain of the speech data (analog/digital) is to put signal extraction at a position in the receiving telephone that gives access to the data in either analog or digital form. Note in particular that the ISDN signal is encoded as A-law.

8.4.1.5 Recording procedure

The recording procedure comprises a whole range of measures, beginning with the calibration of the microphone and ending with the design of proper interaction between the talking subject and the recording manager. A detailed description of various aspects of concern with regard to the recording procedure is presented in Chapter 4.

From a technical point of view, however, the calibration and the positioning of the microphone is of central interest. It goes without saying that calibration is to be omitted in on-site situations like telephone recordings, for example.

8.4.1.6 Recommendations on microphones and speech recordings

For each of the preceding subsections we give a separate paragraph of recommendations:

8.4.1.6.1 Recommendations on microphones

The choice of the right microphone strongly depends on the specific task to be performed. In on-site recording situations, often no decision on the microphone can be made. With respect to the best quality obtainable, however, we can give the following recommendations:

- 1. Always choose professional equipment for speech recordings, especially when it comes to microphones.
- 2. For the flattest possible sensitivity response over the entire frequency range of speech pick a condenser microphone.
- 3. In very quiet environments choose a microphone that exhibits omnidirectional directivity characteristics. They are the easiest to use with regard to position and orientation.
- 4. In reverberant and/or noisy environments a unidirectional cardioid microphone represents a reasonable compromise between noise suppression and flexibility in handling. It eliminates perturbation signals arriving from off-axis directions greater than 65 degrees (-3 db).
- 5. Recordings in a car should be performed with unidirectional microphones presenting a sensitivity response with considerable magnitude attenuation at frequencies below 500 Hz. This is due to the fact that most of the acoustic energy emitted by a car or truck originates from the 100–300 Hz band. Microphones of this type are especially designed for hands-free mobile telephones in cars.

- 316 Part II: Spoken language characterisation
- 6. If a fixed position of the microphone with respect to the speaker's head as well as high ambient noise suppression is crucial, use a headset.

8.4.1.6.2 Recommendations on the recording environment

In order to achieve speech recordings with minimum environmental (room) distortions the following recommendations should be followed:

- 1. If available, recording should take place in an anechoic chamber.
- 2. Placing equipment in the recording room should avoided as far as possible; if it is unavoidable, place equipment as far away from the microphone as possible and, ideally, cover it with acoustically absorbent material.
- 3. A direct feedback path of first-order reflections between mouthmanuscript-microphone should be avoided.
- 4. If negative effects on the talker's prosody, due to missing acoustic room information, are a cause for concern, proper room simulation via headphones is recommended (refer to Section 8.5.2).

8.4.1.6.3 Recommendations on the recording chain: studio

In view of the recording chain we may give the following recommendations:

- 1. As a general guideline, always utilise professional equipment.
- 2. To avoid pop noise have the microphone properly wind shielded.
- 3. If condenser microphones are used, activate the built-in high-pass to suppress potential offsets induced by the phantom-power supply.
- 4. The microphone preamplifier should fulfil studio standards, with a noise figure less than -125 dBu at 200 Ohm, gain range 0–60 db.
- 5. Transmission lines must be properly balanced, properly grounded and shielded, and short.

8.4.1.6.4 Recommendations on the recording chain: on-site

Where the operator has control over components in the recording chain, the recommendations in the preceding section hold true. The field of onsite recordings is quite literally wide open, so that recommendations must be restricted to the very common case of data collection via telephone:

- 1. Do not record speech from the loudspeaker of the receiving telephone. Instead, by-pass the signal at some point prior to the audio stage in the phone.
- 2. If there is a choice, always use digital telephone transmission (such as ISDN) for best possible signal quality. Keep in mind that conditions may not be homogeneous, even within the same network.
- 3. Be aware of certain limits in speech data collected by telephone concerning the obtainable SNR, dynamics and the bandwidth of the signal.

8.4.1.6.5 Recommendations on data collection

- 1. Only utilise professional equipment when it comes to data collection.
- 2. Use digital data storage media, ideally a computer hard disk.
- 3. The standard data format is: sampling rates 16/22.5/32/44.1/48 kHz, 16 bit, linear PCM. Notice the simple conversion ratios for digital signal processing for the sampling rates 16/32/48 kHz and 22.5/44.1 kHz.
- 4. Use computer driven sampling devices. They provide best possible data handling, especially in comparison to DAT recordings.
- 5. Filters involved in the preprocessing of the analog speech signal must be designed to be strictly linear in order to avoid unacceptable phase distortions.
- 6. Digital speech data collected via telephone could be resampled according to the specification of the sampling device. This can often be circumvented by proper positioning of the signal by-pass in the receiving telephone.

8.4.1.6.6 Recommendations on the recording procedure

- 1. Microphone calibration: observe the maximum level a subject produces during a test phase, and set the amplifier gain such that the observed peak level is about 12 db below the maximum possible recording level.
- Microphone positioning (omnidirectional): 20-30 cm distance, 90 degrees incident, 15 degrees off-axis.
- 3. Microphone positioning (unidirectional): 40–50 cm distance, 0–60 degrees incident.
- 4. Microphone positioning (pressure zone): 50–60 cm distance, situated on a table in front of the talker.
- 5. Microphone positioning (headset): ca. 5 cm distance, 20–30 degrees offaxis, at the same level as the lower lips.

8.4.2 Parallel recordings

Speech production is an extremely complex process. It involves several major organs (lungs, larynx, mouth, nose, brain, etc.) and stands in more or less direct dependence to diverse second order physiological parameters such as pulse, blood pressure, or sweat production.

But nobody just talks. There are always facial expressions and gestures involved, which are in fact the only source of information for the interpretation of speech by the deaf. Speech production is clearly highly dependent on biographical factors.

In general it is essential to not only record the pure speech, i.e. the time signal representing the air pressure at the microphone versus time, but to log as much correlated data as possible at the same time.

The range of speech-related parameters to be investigated may be large and often depends on the underlying purpose for which the speech recording is made. Within the framework of this handbook the scope is restricted to the

description of the most commonly used parallel (simultaneous) recording techniques.

However, for obvious reasons, parallel recording must not interfere with actual speech production, either physically or by imposing additional psychological stress on the talking subject.

8.4.2.1 Pitch Determination Instruments (PDIs)

The reasons for determining *pitch* or *fundamental frequency*, in parallel to the time-signal speech data, are numerous. Since pitch determination by machine generally performs a precise and reliable job (Hess 1983), it is frequently utilised for the automatic segmentation and labelling of speech. Second, a laryngogram permits classification of a voice almost at a glance, which is very useful for the classification of talkers according to the recommendations described in Section 8.3. Finally, it might be used to visualise speech for education and rehabilitation purposes.

Although plenty of pitch determination algorithms (PDAs) have been developed in the past decade or so, none has achieved the reliable performance of a PDI. This is even more true of potentially pathological voices.

Although the term *pitch* is often used as a synonym for *fundamental frequency*, the following distinction is sometimes made:

PITCH: perceived speech melody;

FUNDAMENTAL FREQUENCY: the least common multiple of overtone frequencies in a harmonic signal;

PHONATION RATE: the opening/closing rate of the glottis.

1. Mechanical PDIs

The underlying principle of mechanical PDIs is to directly convert vibrations at the throat into an electric signal. The application of these instruments is mainly in education and rehabilitation, e.g. in teaching the intonation of foreign languages, or in the education of the deaf. The mechanical PDI yields an excellent signal for pitch detection; for glottal waveform investigation, however, its output signal is not equally well suited since the detection of the instant of glottal opening and closure is difficult. This is due to inertia effects that mechanical PDIs, which operate on the basis of throat microphones, contact microphones, and accelerometers, suffer from. The most critical point in the realisation of suitable microphones is to decouple background noise and speech of the talker itself from the actual throat vibration. In any case a very tight coupling and extremely good isolation at the throat is needed and these measures may have an adverse effect on the talking subject.

2. Electrical PDIs

Electrical PDIs utilise the change of the electric impedance Z_L of the larynx due to the opening and closing of the glottis. The technical idea is to let the changing impedance of the moving larynx modulate a high-frequency (usually about 1 MHz) circuit (see Figure 8.7).



Figure 8.7: Technical principle of an electrical PDI

The output signal of an electrical PDI is extracted from the HF-voltage by a simple radio-frequency AM-demodulator. Pitch determination from this signal is straightforward, since large jumps at the instant of glottal closure are observed, which may be detected by a threshold analysis on the first derivative of the recorded curve.

Electrical PDIs are commercially available under the names glottograph, laryngograph, or electroglottograph. They are optimal with regard to precision, handling, robustness, and the negligible discomfort they may cause to the talker. In some rare cases the PDI might not work for an individual speaker; when it does work, however, it is fairly foolproof.

3. Ultrasonic PDIs

The principle of this method is based on the fact that the acoustic impedance of air is extremely different from that of flesh, cartilages, and tissue. If a focussed ultrasound beam is transmitted through the vibrating vocal cords, it will only be able to pass if the glottis is closed. If the glottis is open, the ultrasound wave is almost totally reflected due to the impedance mismatch between the tissue and the air in the glottis.

Basically, two different principles of investigating the vocal cords with this method have been developed: the pulse-echo method and the continuous-wave method. To design a PDI, the latter appears most promising. It is based on the idea that the vocal cord vibrations modulate a continuous-wave ultrasound that is transmitted through the larynx at the level of the vocal cords.

Ultrasonic PDIs using continuous-wave ultrasound show an output signal similar to that of electrical PDIs, but unlike the latter have almost 100% amplitude modulation when the beam passes through the vocal cords. On the other hand, the ultrasonic PDI is much more sensitive to vertical positioning of the transducers. That is why we do not recommend this device for parallel recordings, at least not for the untrained operator.

4. Photoelectric PDIs

As with the electrical PDI, the photoelectric PDI is commercially available. It is based on the principle of transillumination of the glottis. A strong light source is placed at the neck below the glottis. Part of the light passes through the skin and the tissue into the trachea. If the glottis is closed, the light is absorbed by the vocal cords, and the pharynx remains dark.

A phototransistor in the pharynx which works as a light transducer picks up the temporal variations of light in the pharynx due to glottal opening and closure. In contrast to the electrical and the ultrasonic PDI, the photoelectric PDI gives a measure of the cross-sectional area of the glottis, not the degree of glottal closure.

The practical application of this technique, i.e. the positioning of the photoelectric transducer in the pharynx, poses a certain difficulty in respect of the long-term consistency of records produced with photoelectric PDIs (not to mention the stress and discomfort this little piece of high-tech might impose on the person it is attached to). However, the short-term performance of this instrument is excellent. It exhibits an exact synchronisation with both the point of glottal opening and glottal closure. Unlike the electrical and the ultrasonic PDI types, measurements with the photoelectric PDI are possible when the glottis does not close completely, e.g. due to voice disease or breathy voice in normal speech.

Accordingly, a photoelectric PDI should be restricted to voice source investigation in basic phonetic and linguistic research in logopedics and phoniatrics; for simultaneous high-quality microphone recordings of speech, however, this technique is not recommended.

8.4.2.2 Physiological measurements

The simultaneous measurement of second order physiological quantities like pulse, blood pressure, body temperature, skin can be useful for some purposes. One might think of taking an EKG and/or EEG parallel to the speech recording. However, the range of possible physiological measurements is large and the choice depends on the specific purpose the speech material is to be made for. Whatever additional recording is decided on, it must again be stressed that any disturbance to the talking subject should be minimised.

8.4.2.3 Mimic and gesture recordings

It is known that speech relates closely to facial expression and gesture; For deaf and dumb persons, direct communication relies largely on lip reading and sign language. In applications which may be related to this domain, we therefore recommend carrying out additional video recordings simultaneously with the speech recordings.

Since sign language takes place in 3-dimensional space and since hands might overlap for some gestures, there should at least be two cameras, separated from each other by a well-defined angle and distance, mounted in front of the talker. For later reconstruction and spheric evaluation, it is crucial to note the exact data related to the relative positioning of the cameras. Commercial systems for 3-D recordings, however, are widely available though not inexpensive.

For lip-reading recordings, a single camera in front of the subject may suffice.

8.4.2.4 Recommendations on parallel recordings

The following recommendations can be given with respect to the use of parallel recording techniques:

- 1. First and most important, none of the additional recording devices should disturb the actual speech production and recording process.
- 2. After consideration of all aspects of selecting the right pitch determination instrument (PDI), we recommend the use of an electrical PDI, also known as glottograph or laryngograph. It is commercially available and affordable, precise and reliable in its function, robust, easy-to-use in operation, and, last but not least, it does not noticeably interfere with the talker's primary task, i.e. to talk.
- 3. For facial expression and gesture recordings choose commercially available 3-D video recording systems where possible.
- 4. Lip-reading recordings necessitate at least a single camera in front of the talker.

For further reading consult Bartlett (1987); Ballou (1987); Davis and Davis (1975); Tomlinson (1990); Hess (1983).

8.5 Environment characteristics

8.5.1 Artificial vs. natural environment

Which environments are to be considered artificial and which natural is a difficult question which will not be discussed within the framework of this section. From the viewpoint of a potential talker, however, the environment starts to be artificial from the moment he notices that a recording is going on. It is generally agreed that knowledge of being taped has a significant influence on speaking style. Consequently, whenever real-life data is needed,

the appropriate recording technique is that of on-site recordings, with all the potential drawbacks with regard to signal quality. In turn, once it has been decided that known influences may be accepted in the interest of complete control over the entire recording procedure, including phonetic and linguistic aspects, the technical setup, and the recording environment, it is strongly recommended that arrange everything be arranged in such a way that the subject is not further disturbed by any undesired phenomena. In other words, in an artificial recording situation such as studio recording one should try to provide the talker with the most natural environment possible.

The environment comprises a whole variety of factors such as lighting, temperature, visual impression and acoustics.

By "natural" we understand "realistic"; for example a situation in which a realistic artificial noise scenario is presented to the talker is considered to be natural. The crucial point is that the talker must not be under the impression that he is to be deceived by some artifice.

Further discussion on the influence of environmental factors on talkers may be found in Section 4.2.4.

8.5.1.1 Recommendations on artificial vs. natural environment

As a rough guide to the conditioning of talkers by environmental factors we give the following general recommendations:

- 1. For some types of real-life speech data on-site it may be preferable not to tell the person to be recorded that he is to be taped (see Section 4.2.2).
- 2. In on-site recording situations, such as in a car, try to simultaneously record any speech related environmental data (e.g. speed of the car, temperature, time of the day, etc.).
- 3. In an artificial recording situation, successful environmental conditioning of a subject can only be achieved if the talker does not have the impression that he is being subjected to unconscious manipulation.
- 4. In artificial recording situations, avoid undesired environmental impact on the subject such as unreasonable room acoustics, insufficient lighting, or unusual low or high temperature. Be aware that this will definitely influence the talker's speaking behaviour.

8.5.2 Auditory environment

It is obvious that for speaking experiments the auditory domain plays by far the most important role of all environmental factors. At this point we assume a studio situation, i.e. a speech recording in an anechoic chamber, and concentrate on how to provide a talker in such a situation with a reasonable impression of the room's characteristics without corrupting the speech signal at the same time. The focus is on simulation of just any room instead of a particular one; the primary goal is to avoid unwanted psychoacoustic effects due to the missing information characteristic of rooms in general. As basic knowledge one has to understand the main acoustic components that influence the perception of one's own voice. These are:

- internal (in-head) transmission of sound inside the head through air filled tubes to the eardrums and through the bones and other tissue directly to the cochlea,
- external (around-the-head) sound transmission through the air around the head to the eardrums, and finally
- images of one's own voice reflected by acoustically relevant surfaces in the environment.

The dilemma is that for the purpose of providing an acoustic signal to the talker without disturbing the actual speech data we need to equip him with *headphones*. But what type of headphone is to be chosen and how can we compensate for its effect with regard to self-perception? How to actually simulate a room, once proper headphone compensation has been installed, and how this gives access to all kinds of acoustic subject conditioning will be discussed in Section 8.5.2.2.

A comprehensive elucidation of various aspects of "Vocal Communication in Virtual Environments" can be found in Lehnert and Giron (1995).

8.5.2.1 Use and compensation of headphones

8.5.2.1.1 Open vs. closed headphones

One could be tempted to say that closed headphones are the ideal solution to the problem of providing a talker with sound. Unfortunately, they are not.

First of all, even hearing protectors which are specially designed to be as "closed" as possible do not provide perfect insulation; especially at low frequencies acoustic shielding is poor. This effect is even more pronounced for closed headphones. The difference between open and closed headphones is a qualitative one and not a principled one, and the terms are somewhat misleading.

A closed headphone will change the radiation impedance at the end of the auditory channel towards the free sound field and therefore the in-head sound transmission mechanism of the voice will be severely affected. Equalisation of this effect, especially at higher frequencies, is very difficult to achieve.

Another effect caused by modified radiation impedance is that sensitivity to hearing one's own blood flow is increased. Also, sounds that are at least partly perceived by the body, e.g. one's own footsteps, are greatly modified and in general sensitivity to structure-transmitted sounds is increased.

The only benefit of using closed headphones is that external sound is shielded somewhat better than with open headphones. but in a rather

quiet environment, such as an anechoic chamber, this should in any case not be of major concern. In the other direction, i.e. from the headphone to the microphone, it is necessary to be aware of the possibility of undesirable feedback (cf. Section 8.5.2.1.4).

It is always recommendable to use open headphones for the sound conditioning of a talker.

8.5.2.1.2 Using diffuse-field equalised headphones

Sound reproduction can be performed in two ways, namely by either using loudspeakers or by using headphones. Only headphones offer a means of reproducing sound in a systematic and well-defined manner. However, for that purpose a specific headphone or a specific kind of headphone must be chosen.

1. Equalising headphones

Every headphone may be described by its frequency response. The frequency response of a headphone can be measured by applying the headphone to an ear-like device, for instance a dummy head with a simulated auditory channel and an ear coupler which has the same mechanical characteristics as the human ear drum, middle ear and the inner ear. Then, for all frequencies in the audible range, the ratio of the Fourier spectrum of the sound pressure in front of the (artificial) ear drum to that of the headphone voltage can be measured. This (rather complicated) method permits prediction of the behaviour of a headphone when it emits sound to a human ear. But what should the frequency response of a headphone look like?

The easiest method of equalising a headphone would be to equip it with a flat frequency response. But this would result in a very unnatural sound. The reason for this is that the human outer ear, the head and the torso are a direction-dependent filter for incoming sounds. That means that the spectral shape of a sound event changes depending on its direction. Our brain is able to do inverse filtering with respect to the sound source position. The result is that when a sound source rotates around a listener's head the perceived timbre of the sound remains more or less unchanged, although the spectrum of the sound measured at the ear drum changes dramatically. The consequence of this effect is that a headphone can only be equalised correctly for a particular condition of sound incidence.

Recently, two equalisation techniques have been developed. The first one is the so-called *Free-Field Equalisation*. A free-field equalised headphone produces the same spectral distribution of sound at the ear drum of the listener as does an ideal loudspeaker placed under free-field conditions (e.g. comparable with an anechoic chamber) in front of the listener.

The second one is the so-called *Diffuse-Field Equalisation*. A diffuse-field equalised headphone, when fed with white noise, produces the same spectral distribution of sound at the ear drum of the listener as appears in a diffuse field. In a diffuse sound field the direction of incidence is evenly distributed over all directions (e.g. in a reverberation chamber).

2. Selection of the right equalisation

The main difference between free-field and diffuse-field equalisation is that a free-field equalised headphone is equalised with respect to the forward direction, whereas the diffuse-field equalised headphone is equalised with respect to an average over all directions of incidence.

Most of the sound signals that are to be reproduced via headphones consist of incoming sounds from various directions. Such sound signals would require the use of a diffuse-field equalised headphone, because this type is the better choice in the sense of "least mean error". Even for sound sources coming from a single direction the diffuse-field equalised headphone is the better choice when the direction of incidence is not close to the forward direction.

Another point is that the free-field equalisation function varies from person to person much more than the diffuse-field equalisation function does. So an averaged diffuse-field equalisation function is valid for more people.

Both the diffuse-field equalisation function and the free-field equalisation function have been standardised, but so far, there are still differences among diffuse-field equalised and free-field equalised headphones made by different manufacturers.

The most important reason for choosing diffuse-field equalised headphones is that recordings made for diffuse-field equalised headphones also yield good results when reproduced via loudspeakers and vice versa. Using diffuse-field equalised headphones offers the best compatibility to common recording techniques, at least in the opinion of many experts.

3. Using dummy heads and binaural simulation techniques together with headphones

When reproducing dummy head recordings or recordings made by using binaural simulation techniques via headphones, localisation errors and coloration can only be avoided when the headphones fit well together with the recording technique. In the case of dummy head recordings, this simply means the use of free-field equalised dummy heads with free-field equalised dummy heads and the use of diffuse-field dummy heads together with diffuse-field equalised headphones. Diffuse-field equalised dummy heads (e.g. the Neumann KU 81) are mainly used for listening purposes whereas free-field equalised dummy heads (e.g. the "Aachener Kopf", the "Aachen

Head", made by the Head Acoustic company) are mainly used for measurement purposes.

When making recordings using binaural simulation techniques, the catalogue of outer-ear transfer functions should also be diffuse-field equalised.

4. Various diffuse-field headphones

At the moment, diffuse-field equalised headphones are available from different companies, for example:

- Stax SR Professional Lambda incl. diffuse-field equaliser
- AKG 240 DF
- Sennheiser HD 250 linear
- Sennheiser HD 540 reference

The Stax is an electrostatic headphone and it is delivered together with a preamplifier which contains the diffuse-field equalisation. It is probably the best headphone available on the market, but it is very expensive. All headphones are open except of Sennheiser's HD 250 which is a closed type. The company "Beyer" sells a headphone which is delivered together with a passive network for diffuse-field equalisation.

In the SAM Project, the AKG 240 DF was selected as a low-cost standard headphone, while the Stax SR with diffuse-field equaliser is a good choice as a highest-quality reference system.

8.5.2.1.3 Insertion-loss compensation

Since the subject wears a headphone, the transmission of the speech sound through the air around the head is significantly disturbed. In order to model natural self-perception of the voice, this effect has to be compensated for. The sound field outside the head can be considered as a linear area that is free of sound sources. In such an area the sound pressure signal at any point can be reconstructed from the sound pressure signal at any other point just by knowing the correct transfer function between both points. The task of measuring this transfer function is similar to the well-known procedure of determining a Head Transfer Function: miniature microphones are placed at the entrance of the blocked auditory channels. While speaking, the sound pressure signals at the reference point (recording microphone) and the ear microphone are recorded simultaneously. The magnitude of the transfer function may be obtained by averaging the energy of the short term spectra of both signals and dividing the resulting values at the ear drum by those measured at the reference point. During this procedure the phase information is lost. A plausible phase can be generated by calculating the minimum phase function. The same procedure has to be done with the subject wearing the headphone such that the resulting compensation function is given to 1 - l(f), where l(f) is the result of complex valued division of both transfer functions.

A sensitive matter is the choice of the reference point; if authentic compensation is desired, this point, i.e. the recording microphone, must be located as close to the mouth as possible and it must not move significantly during a recording session. The use of high-quality headsets is therefore strongly recommended for that purpose.

Since the level of compensation is critical, the insertion-loss compensation function has to be determined for each talker individually.

A rather practical approach to the problem of insertion-loss compensation is to ask the subject to just turn the gain of his own speech signal – equalised with the Head Transfer Functions of the ear-to-mouth direction and fed back to the headphone – up to the point where it sounds as "normal" as possible.

8.5.2.1.4 Feedback compensation

Feedback compensation, i.e. the compensation for possible sound transmission from the headphones to the recording microphone, is not considered to be necessary.

If pure headphone compensation and room simulation is requested, sound is only emitted from the headphones simultaneously with the speech and thus would not degrade the signal-to-noise ratio of the recorded signal.

Commonly used background sound, such as concurrent speakers or an underlying noise floor that might be intended for subject conditioning should also not jeopardise the speech quality. A sensitive upper limit for those conditioning signals is considered to be a sound pressure level of 85 db. Above this limit, audible feedback from the headphone to the recording

microphone should be taken into account during non-speaking intervals.

8.5.2.2 Modelling of the auditory perspective

Spatial auralisation of sources can be performed by real-time filtering of sound signals with the Head Transfer Function of the talker in combination with a modelling system that calculates the spatial map of secondary sources and the corresponding filter functions. In a successive stage the contributions of all secondary sources are filtered with the Head Transfer Functions for the corresponding directions of incidence. Figure 8.8 displays the auditory subsystem of the so-called SCAT-LAB that has been developed in the course of the ESPRIT basic research project 6358 SCATIS (Spatially Coordinated Auditory/Tactile Interactive Scenario). Since SCATIS was originally designed for unidirectional simulation (passive subject) it has been augmented according to the needs for speech recordings in anechoic chambers, i.e. microphone feedback to the DSP-network has been established,

and the database has been expanded by the insertion-loss compensation function of the headphone.



Figure 8.8: Augmented auditory subsystem of the SCATIS VE generator

8.5.2.2.1 Talker mirroring in the virtual room

With respect to the modelling of room acoustics, the subject's voice is a sound source like any other and may be modelled as such, though with a few exceptions:

- 1. The direct sound component is to be handled differently. In the original room acoustics model, the sound source and the receiver are located at the same point, which is not permissible for the talker mirroring. Here, the direct sound component is produced by the speaker himself and reconstructed correctly, thanks to insertion-loss compensation described above. The consequence for room acoustic modelling is that reflections have to be rendered as normal whereas direct sound has to be omitted.
- 2. The modelling of directivity with linear direction-dependent filters implies that the source or reflections is located in the far field of the source. Unfortunately, this is not the case since the point where the speech signal is picked up is very close to the mouth and therefore definitely in the near field range. Due to the linear character of the sound field, this might be

corrected by a single linear filter for all directions of emission. This filter is given by the ratio of the sound pressure spectra measured at a point in the reference direction which is sufficiently far away from the sound source.

3. With the SCAT-LAB a head-tracker is mounted to the subject which provides the system with all the necessary information to dynamically retune the position dependent parameters of the room simulation. Therefore the talker is free to move around within a recording session. Only the microphone position has to be well defined and fixed during the recording.

8.5.2.2.2 Subject conditioning

Up to this point we have described how to properly compensate for disturbing acoustic effects due to insufficient self-perception resulting from the headphone, or due to missing sound reflections in an anechoic chamber. In fact the technical setup for this also allows for a virtually unlimited range of acoustic conditioning of the subject. This includes ordinary noise of defined level and spectra, or simple monaural interaction between the talker and the recording manager, as well as complex scenarios such as dialogues in the entrance hall of a railway station with incoming trains, heavy reverberations, and concurrent speakers from different directions. Additional sounds such as these may come from a tape or they may be played in on-line. In Figure 8.8 this is summarised in the block labelled "Audio Sources". For later mixing and scenario analysis it is advisable to synchronise the recorded speech with the background signal.

8.5.2.3 Recommendations on the auditory environment

The following recommendations are given for speech recordings in a very quiet recording environment (e.g. an anechoic chamber) that deprives the talking subject of most of its "natural" room impression. The goal is to compensate for this deficit and to decouple any acoustic feedback to the talker from the actual speech recording at the same time:

- 1. Use headphones to provide the talker with acoustic feedback/background/interaction/control in order to avoid interference with the speech recording.
- 2. Choose so-called open headphones rather than closed types. They yield less degradation of acoustic self-perception and minimise weight related discomfort.
- 3. Install proper insertion-loss compensation for the headphones. Use highquality headsets for that purpose.
- 4. For acoustic stimulation from directions other than the front choose diffusefield equalised headphones.
- 5. When headphones are used in combination with binaural simulation techniques make sure both are equalised in the same way (diffuse-field or freefield).

6. For proper acoustic room simulation make use of existing systems, such as the SCAT-LAB developed under ESPRIT project No. 6358.

8.5.3 Visual environment

The visual environment clearly starts with the lighting. In general, this should be kept at a level such that the talker feels comfortable and "normal". This should be mentioned because artificial recording environments are often illuminated at either very bright or rather dark levels.

Apart from this most basic property of the visual environment, additional devices are available to model the visual environment of the talker.

8.5.3.1 Head mounted visual displays ("eyephones")

In most of today's practical implementations the visual displays of so-called Virtual Environment (VE) systems consist of two LCD monitors and some optics which are located close to each eye. The whole setup is assembled in a kind of "helmet" to be worn by the subject. The helmet also usually incorporates the head-tracker as well as the headphones.

From the acoustic room simulation point of view, wearing a helmet has a number of specific disadvantages and advantages, the first of which might be technically compensated for.

With regard to the paradigm described above, however, in which that a talker should not be aware of being subject to some kind of unconscious manipulation, it is obvious that a helmet of this type is cannot provide anybody with anything like a "natural" impression of the visual environment. Wearing a device as bulky and weighty as a helmet, no one can realistically be expected to behave in a natural way.

Finally, so far no system is known that would be technically capable of providing a realistic simulation of a natural environment, even if money is no object. In this respect, today's visual VE-systems are not considered to be superior to large scale frontal visualisation.

8.5.3.2 Frontal visualisation

Frontal visualisation is a quite common technique in the speech recording area. Its application ranges from ordinary text prompting to more or less realistic face to face interaction between the talker and the recording operator.

When using frontal visualisations, especially those of rather large format, one should be aware of the fact that they cause acoustic reflections which in turn degrade the recording quality. Consequently, we would not recommend trying to project anybody into a "virtual" visual environment by seating him in front of a very large scale screen, if reflections might be a source of concern. The use of a screen should be restricted to pure prompting applications in which the monitor can be kept relatively small, and situated sufficiently far away from the microphone.

8.5.3.3 Recommendations on the visual environment

- 1. Keep the lighting of the recording environment at a "normal" level; most recording rooms appear to be either too bright or too dark.
- 2. State-of-the-art eyephones (head mounted visual displays) are not considered to be suitable for providing a realistic or natural visual environment.
- 3. In very quiet environments, such as anechoic chambers, the size of potential prompting screens should be kept as small as possible and away from the microphone.

8.5.4 Miscellaneous environmental factors

The most important environmental factors apart from acoustics and vision are possibly temperature, force impact (movement), as well as taste and smell.

As with the eyephones, no device is known up to this point that would be capable of simulating any of the above in a natural way, though some development in this direction is going on. By "simulation in a natural way", we mean in such a way that a subject who felt a simulated touch by a virtual object, for instance, would actually behave as if he really had been touched.

For further reading consult Lehnert and Giron (1995).

8.6 Recording chain characteristics

The recording chain, set up by various electroacoustic devices like microphones, amplifiers or digital filters, may be seen as a single "black box" with a single input and single output. This is commonly called a *system*. A recording channel is a transmission channel between the signal to be recorded and the recording medium (i.e. analogue or digital tape recorder, hard disk). In the cases of one, two and more channels we speak of monophonic, stereophonic and multi-channel recordings, respectively.

The effects of using arbitrary recording system are twofold. First, the signal sent from the input through the system to its output is altered compared to the intended, but ideal, transformation in amplitude and/or delay time. Since ideal systems are to be found in theory only, unwanted side effects (distortions) due to a deviation of system behaviour from the ideal specification have to be considered.

The second kind of system characteristics are discussed below. Although these characteristics are not desired, it is important not to neglect them, either when choosing a specific device or in reporting a system configuration. For that reason, all manuals of the system components under consideration should be checked for the relevant distortion data.
332 Part II: Spoken language characterisation

8.6.1 Signals and systems

A physical process that shows the same state after certain time epochs is called an oscillation. We can distinguish oscillations with identical and different sections, called periodic and aperiodic oscillations, respectively.

Both kinds of vibration are signals that may be observed in linear as well as in non-linear systems. A linear system is understood as a system with only linear electrical and mechanical components. In such a system the ratio between the input and output amplitude is always constant, in contrast to non-linear systems. The superposition principle that holds for linear systems says that an oscillation occuring in a linear system is not influenced by cooccurring oscillations. Most important, this includes also the frequency concerned. This principle is especially important for a speech recording chain, where many oscillations with different frequencies always exist at the same instant.

A simple but more mathematical interpretation of superposition may be formulated as follows. The superposition principle requires that the response of a system to a weighted sum of signals is equal to the corresponding weighted sum of outputs of the system to each of the individual input signals. This includes the additive and multiplicative (scaling) properties of linear systems. In other words, a relaxed, linear system with zero input produces a zero output. If a system produces a non-zero output with a zero input, the system may be either non-relaxed or non-linear. If a relaxed system does not satisfy the superposition principle, it is called non-linear.

Systems can also be classified into the two broad categories of time-variant and time-invariant systems. A system is called time-invariant if its input– output characteristics do not change with time. On the other hand, if the outputs to the same input differ for the same system at different times, the system is called time-variant. A common class of systems is that of Linear and Time-Invariant (LTI) systems.

An important property that must be considered in any practical application of a system is *stability*. A system is defined to be bounded-input-bounded output-stable, if and only if every bounded input produces a bounded output. An LTI-system is stable, if the output response to an input impulse is absolutely summable. Unstable systems usually exhibit erratic and extreme behaviour and cause "overflow" in any practical implementation.

The devices of a recording chain are not LTI-systems in the ideal sense, and consequently we have to reckon with non-linearities. The oscillations at certain frequencies come under mutual influence, and new frequencies are produced. If this process was intended, it is called modulation or demodulation. Otherwise, we call it a distortion.

In the following subsections we consider signal behaviour during transmission through linear and non-linear systems. We start with simple sinusoidal oscillations.

8.6.2 Linear distortions

Any frequency spectrum that passes through a linear system can undergo only linear distortions. In practice these distortions may emerge by a change in signal amplitude and/or a change in signal phase for each frequency component. Consequently the distortions of linear systems are divided into two categories: amplitude and phase distortions.

8.6.2.1 Amplitude distortions

The characteristic of this kind of distortion is a frequency-dependent change in the output or response amplitude a_o to input or excitation amplitude a_i ratio, the so-called *transmission factor*:

$$f = \frac{a_o}{a_i}$$
, or transmission measure $m = 20 \log \frac{a_o}{a_i}$ (8.5)

For an acoustic or mechanical output quantity the transmission factor is also called *efficiency*, while it is called *sensitivity* in the electrical case.

The dependency $f = f(\omega)$ or $m = f(\omega)$ is called the *frequency response*, and may be obtained by calculation or by measurement. A graphic representation of the *amplitude-frequency curve* is called a *Bode diagram*. The transmission factor at a test frequency f_t compared to the transmission factor at a reference frequency f_r (commonly 1000 Hz for audio devices) is called damping distortion. In case of the same input amplitude for the test and reference frequency, linear and logarithmic damping distortion is defined respectively by:

$$f_d = \frac{a_{otest}}{a_{oref}}$$
 and $m_d = 20 \log \frac{a_{otest}}{a_{oref}}$ (8.6)

For instance, a system may be described by "frequency range 40 to 10000 Hz ± 1 db". This means that at all frequencies within the cited frequency range B the damping distortions are lower than ± 1 db. If a cited frequency range goes without a specified damping distortion, we can assume a maximum distortion of -3 db compared to 1000 Hz, corresponding with a maximum decrease in the transmission factor by $\frac{1}{\sqrt{2}}$. Where at constant input amplitude the output amplitude of a system changes in proportion to the frequency, the damping distortion is specified by "dB per octave".

In general terms, B is the bandwidth with an approximately horizontal amplitude-frequency response and may also be specified by the lower and upper cut-off frequency. Below and above these frequencies the response differs appreciably from the response within B. We should bear in mind that a specification of a lower and an upper cut-off frequency gives no information either on the strength or on the kind of distortion (dips/notches or peaks) below and above B. Additionally, the quality of linearity within B is not

334 Part II: Spoken language characterisation

fully specified. Therefore a more practicable description of the amplitude distortion may be obtained from a tolerance mask that shows the amplitude–frequency range containing the frequency response.

The frequency range required for high listening quality is to a large extent application dependent. High-quality reproduction of music may be obtained within a range from 40 to 15000 Hz, whereas a range from 20 to 20000 Hz is recommended for professional audio applications. In case of speech signals the frequency range varies between 300 to 3400 Hz (telephone quality), 70 to 8000 Hz (reasonable quality) and 40 to 15000 Hz (high quality). A sharp reduction of the frequency range of speech that can be audibly detected by 80% of the subjects occurs for ranges of 120 to 7900 Hz (male speaker) and 220 Hz to 105000 Hz (female speaker) (Webers 1985).

In the case of digital audio, attention should be paid to the choice of the sampling rate, which must be equal to or higher than twice the upper cut-off frequency. Therefore a compromise has to be met between better listening quality obtained from higher frequency energy on the one hand, and higher storage and processing demands on the other. For professional applications of speech or application-independent "flawless" recordings, we recommend a sampling rate of 32000, 44100 or 48000 Hz. This ensures high quality as well as flexible conversion of the data to standardised digital audio formats. The perceived degradation by irregularities in the amplitude-frequency response are dependent on the form of the deviation from the horizontal "ideal" curve. It has been shown by listening tests that distortion peaks are much more disturbing than distortion valleys or dips: at 90 Hz a peak of 10 db was detected in 40% of the test cases, and a dip of 25 db in only 30% (Rossi 1988). This holds for music, white noise and speech, and the signal presentation was done via headphones in this evaluation. Consequently, dips are considered acceptable, whereas distortion peaks are perceived as very unpleasant. Additionally, distortions at high or medium frequencies are more annoying than the same distortions at low frequencies (Eargle 1976). These facts, which can also be corroborated by speech intelligibility tests, have been taken into consideration in standards for high-fidelity electro-acoustic equipment.

8.6.2.2 Phase distortions

So far we have considered one conformity condition of the recording chain: a constant or frequency independent gain G. Therefore a theoretical transfer of an impulse over the recording chain may lead to an output impulse that is a scaled version of the excitation. Additionally, the output impulse may be delayed in time by a frequency independent delay τ , corresponding with a linear phase response in the frequency domain.

If a speech spectrum is cut into two components below and above a cutting frequency, the smallest audible delays between these components are ap-

proximately 10 ms and are to be found for cutting frequencies between 500 and 2000 Hz (Webers 1985). The smallest perceivable delays are speaker-dependent and increase in case of a reverberant environment.

Today the effects of linear phase distortions on listeners are still under discussion. There are no defined requirements with respect to this kind of linear distortion.

8.6.3 Non-linear distortions

Any non-linear component in the recording chain may introduce *non-linear distortions*. The characteristic of this kind of distortion is the creation of new frequency components in the spectrum, which may be non-harmonic in the worst case. Depending on the generation process of the new frequencies, harmonic distortions, intermodulation distortions and difference-frequency distortions are to be differentiated.

A system component shows a non-linear behaviour if the output amplitude is not proportional to the amplitude of the input signal. In other words, the output-to-input amplitude function is not a straight line but an arbitrary curve. Typical non-linear devices are limiters or noise gates. Every overload (i.e. when the signal exceeds the maximum dynamic range), including amplitude clipping and frequency aliasing, produces non-linear distortions. The resulting degraded speech signals are judged to be unclear, rough or croaky.

Due to the manifold factors influencing the detectability of non-linear distortions by human listeners (for instance loudness level and frequency components of the distorted signal), values for just noticeable distortions cannot be provided. However, the degree of intermodulation and differencefrequency distortions indicate the degree of degradation due to a nonlinearity. Accordingly, for professional applications with high fidelity the intermodulation-distortion factor should be less than 1%, since 0.2% intermodulation distortion may already lead to audible effects.

8.6.4 Recommendations on recording chain characteristics

The following recommendations may be given with respect to the design of the recording chain:

- 1. Choose a recording equipment that ensures a flat amplitude frequency response within the required frequency range.
- 2. Try to avoid components that may introduce linear distortions in the form of spectral peaks, especially at medium and high frequencies.
- 3. Avoid any overload or aliasing of a recorded signal since this introduces critical non-linear distortions. This has to be emphasised for digital recordings due to the sharp clipping of all information above the most significant bit.

For further reading consult Rossi (1988); Webers (1985); Eargle (1976).

336 Part II: Spoken language characterisation

8.7 Reproducibility assurance procedures

Reproducibility means the ability to reproduce a signal under the same conditions. The requirement of ensuring the reproducibility of speech recordings can, above all, be motivated by the following factors:

- For speech database production:
 - to enable adequate data selection for a specific task (i.e. for evaluation of speech recognisers) from a huge pool of different speech corpora, based on a complete specification of the speech recordings, and
 - to enable future extensions of the database (i.e. with different speakers or languages).
- For the production of speech data to be exploited as test stimuli in speech assessment procedures:
 - to ensure that the tests may be repeatable with similar data for reasons of extension or verification,
 - to ensure that the test results may be interpretable correctly, considering all information about source and channel characteristics, and
 - to allow making comparisons with experimental results gained from other test procedures.

These two goals of common database production and test stimuli generation are not further differentiated. While in Section 8.7.1 some features are suggested to describe the recording conditions, standardised methods and data are proposed in Section 8.7.2.

8.7.1 Description of recording conditions

The characterisation of the signal source (the speaker), of the transducers, and of the environment is presented above in sections 8.3, 8.4 and 8.5. Discussion of these features will be complemented by focussing here on the signal itself, which may be described by temporal or spectral parameters. Obviously, these signal parameters have to be elucidated and supplemented by further information concerning every step of signal generation: a certain signal characteristic may be determined by the instructions that were given to the speaker before reading, by the speaker's voice characteristics, by the recording environment and/or by any electro-acoustic component contained in the recording chain. In order to answer such questions after the recordings have been produced, information on every part of the recording chain has to be collected carefully. Refer also to Appendix D for speech recording protocols, and to Appendix C for file formats used in SAM.

The features mentioned in the following selective list may give an impression of the kind of data that have to be collected in order to ensure reproducibility:

- environment characteristics (simple in the case of an anechoic room);
- scripting: format of the manuscript, carrier context per recording item, additional introductory or ending texts;
- speaker setup: briefing and reading instructions, positioning of the speaker, posture of printed script (reflections!) or monitor scripting, microphone distance and direction;
- speaking style characteristics: speaking rate (syllables per second), average fundamental frequency, intensity, etc.;
- mic type and settings (if available): directional characteristics, LP-/HPfilters, use of a wind protection etc.;
- type and settings of all used electro-acoustic components (preamplification, analog-to-digital conversion, recording level, noise reduction, etc.);
- overall bandwidth by lower and upper cut-off frequency;
- data storage media (tapes);
- digital signal representation: sampling rate and word length in case of linear PCM, or coding method and parameters, additional information on speech files specified by supplementary files or file headers (i.e. Windows *.wav format; refer also to Appendix C for speech file formats);
- post-processing of the speech signals: conversion of data formats, preemphasis, special effects, etc.

8.7.2 References and standards

Another way of supporting reproducibility can be achieved by the application of standardised signals or channels. This increases the chance of setting a common basis in signal generation techniques that are applicable also by audio non-professionals. Additionally, the variety of the available speech databases can be decreased.

8.7.2.1 Reference signals

The primary way to take advantage of standardised signals is to rely on speech signals that have been previously produced and collected in a spoken language corpus (cf. Chapters 3 to 5 on Spoken Language Corpora). A description of public domain spoken language corpora is given in Appendix L; for an up-to-date overview over available corpora and their distribution we recommend contacting the relevant speech agencies like ESCA, ELRA or LDC via the Internet.

If special signals are needed for measurement purposes that are not already available, a standardised scheme for signal generation should be considered. Various signal generators exist in hardware and software that produce for instance sinusoidal signals at an adjustable frequency or a wide or narrowband noise. Noise with constant power density over constant-bandwidth intervals is called *white noise*, and with constant power per third or octave (constant relative bandwidth) *pink noise*. Another common reference signal is the artificial voice that is made up of a sequence with standardised

338 Part II: Spoken language characterisation

glottal pulses reflecting typical long-term frequency characteristics of speech (CCITT 1988a).

For the measurement of sound intensity, for instance the intensity of environmental noise, a sound level meter has to be used. This instrument yields the sound level in db, but includes an optional weighting curve over frequency. Three weighting curves have been defined, referred to as A, B and C, that more or less correspond to the equal-loudness contours of the human ear at three different sound levels (cf. Section 8.2.3 about isophones). In practice, only the A-weighting is used, yielding an A-weighted sound pressure level denoted as L_A in dB(A).

8.7.2.2 Reference channels

Suitable reference signals have to be chosen, but even so, the channel characteristics must not be neglected. To ensure a clearly defined input to the channel, an anechoic room is recommended for positioning the sound source (the speaker). Furthermore, the acoustic characteristics of a human speaker may be simulated by an artificial mouth, and the human receiver (head with two ears) may be simulated by an artificial head (the ear) (Blauert 1983). Both influences can also be emulated by signal processing techniques (cf. Section 8.8).

The fact often neglected that the behaviour of electro-acoustic transducers like microphones is not constant over a long period. Therefore these channel components have to be calibrated at least at the recommended intervals. Check the data sheet or ask the manufacturer of the transducer for calibration time intervals and methods.

8.7.3 Recommendations on reproducibility

We summarise the recommendations about reproducibility as follows:

- 1. Information on every part of the recording chain has to be collected as carefully as possible, with characteristics of the speaker (cf. Section 8.3), transducers (cf. Section 8.4), the recording environment (cf. Section 8.5) and of the recording channel.
- 2. For choosing speech signals, try to apply reference signals from standardised Spoken Language corpora as far as possible.
- 3. If there are no appropriate signals available, try to rely on standardised signal generation schemes and on reference channels as far as possible.
- 4. Check the electro-acoustic transducers (microphones) for requirements and methods of calibration.

For further reading consult CCITT (1988b); Blauert (1983).

8.8 Analysis and conditioning tools for speech signals

This section introduces some common tools for the analysis and modification of speech signals. Important standards and references are mentioned for the largest field of applications, namely computer-based, digital processing of speech signals. Again, we call upon the reader to look for standards for ensuring compatibility between computer systems and languages as well as to benefit from the available technology as much as possible.

8.8.1 Signal analysis and measurement

8.8.1.1 Hardware devices

Due to the fact that most signals are already recorded or automatically generated into a digital representation and stored in a machine-readable way, the use of hardware devices for speech analysis and measurement has more and more receded into the background. The most wide-spread tools are

- a sound level meter for measuring the weighted sound intensity (in dB),
- and a spectral analyser for measuring the energy distribution over frequency.

The use of hardware devices is in many cases motivated by mobile use of long-period and on-line sound measurements.

8.8.1.2 Software tools

A wide selection of software tools is available in the field of signal processing, some of which are specially equipped with functions for speech processing. Included among these functions are algorithms for spectral analysis, formant analysis, F0 determination, speech coding and digital filtering up to complete packages for speech synthesis or speech recognition. The functionality, efficiency and cost of these tools differ widely, depending on the family of applied computers and operation systems. Many tools are general purpose, and constitute a complete speech processing toolkit. Most of the more elaborate systems include acoustic playback, graphical data representation, data import and export functions, history tracking, and programming interfaces with ready-to-use function libraries.

Some software packages need the support of special hardware like digital signal processors, but there exist also cheap and easy-to-use programs running on standard Windows-PCs (like "CoolEdit", a shareware tool that can be obtained via the Internet). Speech processing packages include ESPS/waves+, matlab, SLM and RESAM.

340 Part II: Spoken language characterisation

8.8.2 Signal conditioning

Signal conditioning includes all signal modifications aimed at optimising signal characteristics for special purposes. Examples of such purposes are presented in the next sections.

8.8.2.1 Modification of speech parameters

Modification of speech signals may intentionally influence some of the following parameters:

- The amplitude-frequency spectrum of a speech signal can be altered by digital filters such as low-pass, high-pass or other types of filter. Other methods modify the speech characteristics parametrically: the most common coding scheme describes the strength and frequency of the time-variant vocal tract resonances called formants.
- Prosodic characteristics like speaking rate and fundamental frequency may be modified independently in the time domain by applying a PSOLA algorithm (Moulines and Charpentier 1990). Most PSOLA algorithms need speech signals labelled with pitch marks (fundamental period pointers).
- The sampling rate can be converted by using an algorithm for sample frequency conversion. In decreasing the sampling rate the upper cut-off frequency and signal bandwidth also have to be adopted.
- Other signal modification procedures modify not primarily features of the speech signal itself but channel or environmental characteristics and are mentioned below.

8.8.2.2 Signal contamination

Signal contamination aims at generating "non-flawless" speech signals under well-defined conditions. Starting with ideal "flawless" speech, a wide range of signal and channel characteristics is subsequently adjustable by controlled speech contamination. The "flaws" can be chosen from a variety of linear or non-linear distortions (cf. Section 8.6). Possible applications are the stimuli generation process for the evaluation of the robustness of speech recognisers or the assessment of speech coding algorithms under non-ideal conditions.

8.8.2.3 Binaural processing

Binaural processing simulates the signal processing scheme of the human auditory system (Blauert 1983). By investigating the relevant signal parameters from both ears it is possible to make full use of signal features that are impossible to obtain from monophonic signals. These features include positions, movements and extensions of sound sources. In the field of speech, interesting applications for binaural techniques are speech enhancement and voice separation, for example in connection with the so-called *cocktail party effect*, i.e. the ability to track one voice among many.

8.8.2.4 Simulation of environmental factors

Special algorithms like room simulation or reverberation tools allow subsequent introduction of acoustic environmental characteristics into the speech signal (cf. Section 8.5 on environment characteristics). Beyond impressions like room size and wall properties, sophisticated virtual auditory environments can be simulated in combination with binaural processing techniques. Although the speech signal has been produced independently of these environmental factors, arbitrary room acoustics can be created from a single recording and simulated without the physical existence of the room.

8.8.3 Recommendations on analysis and conditioning tools for speech signals

The following recommendations may be given with respect to the use of analysis and conditioning tools for speech signal processing:

- 1. Look for standards to ensure compatibility between computer systems and languages.
- 2. Keep a record of each step of signal analysis and modification.
- 3. If possible, use auralisation and visualisation functions to check the results of each signal processing step.

For further reading consult Rabiner and Schafer (1978); Moulines and Charpentier (1990); Blauert (1983); Oppenheim (1978); Atal et al. (1991).

342 Part II: Spoken language characterisation

Part III

Spoken language system assessment

9 Assessment methodologies and experimental design

9.1 Introduction

9.1.1 How to read this chapter

This chapter is about methodology for assessing various components involved in language engineering: how to go about sampling enough speakers to ensure that you can make claims about how likely the results are to generalise to a speaker population at large (where population refers to your target market and will vary from application to application); how to compare performance of your recogniser or synthesiser with others that are on the market; how many speakers to include in benchmark tests of speaker verification systems to appraise performance, and so on. For these purposes, an understanding of how to analyse your data statistically is needed.

At other times a user might need to test some very specific idea about, for example, what is going on in his recogniser, whether some gambit for mimicking other people's voices will allow impostors to break into a speaker verification device, what the critical acoustic attributes are that govern the perceptibility of a message in order to improve the systems and how to set up experiments with dialogue systems to check whether they will work adequately for some purpose before committing design engineers to their implementation. The way of approaching the latter group of questions calls for an understanding of the steps involved in setting up and analysing experiments.

The information provided is, then, going to cover general techniques from many diverse areas both in terms of techniques (statistics and experimentation) and applications (including the above examples and many more). Therefore, this chapter cannot hope to be exhaustive in terms of its coverage nor choose an example for assessment which is directly applicable to all needs. However, though there will not be an example for every application encountered, the methodological tools provided should offer some idea of the way to approach many problems that will be encountered. The particular examples for illustration have been chosen in consultation with authors of some of the other chapters. The chapter can be expected to provide information on the following points.

- 1. What will not be presented here are statistical analyses of, for example, the statistical corpora described in other chapters. What is presented here is some of the background that will allow access to the ideas and literature appropriate for tackling the analyses themselves.
- 2. Statistical development, experimental techniques and engineering products and techniques are advancing at a rapid pace. However, statistical and

experimental know-how has not featured to any great extent in language engineering, and statisticians and experimentalists usually have not drawn on examples or considered the engineers' concerns. Thus, many of the "recommendations" made here are a first attempt to tackle these issues. There are often many ways of achieving a particular goal and the limited number of options that it is possible to consider can only give a narrow perspective. As these ideas are tried out, other preferred alternatives will undoubtedly arise. Thus, at least some of the recommendations are likely to be short-lived.

- 3. It has to be assumed that some of the readers of this handbook have had practically no previous experience in the formal methods of statistical analysis. For this reason, it is necessary to cover basic background in statistics in some detail. On the other hand, authors of the other chapters have raised questions about how to address questions statistically which call for advanced techniques. In a chapter of this size, it is not possible to cover both or, to some extent, either topic comprehensively (even introductory texts in statistics usually run to 400 pages). In the text, we have attempted to cater for the needs of both groups: For those with no statistical background a swift overview of the basics is given with illustrations of how these techniques apply to language engineering problems. The more advanced topics are dealt with by pointing out when a topic may be appropriate and the steps to go through. Since those individuals who will want to use these more advanced techniques usually already have some understanding of statistics, at this stage they will have to go to one of the texts referred to for dealing with the actual computational steps.
- 4. Experimentation also raises problems of scope, depth and rigour: For instance, it would be straightforward to describe phonemic labelling of a synthetic speech continuum. This might include describing a speech continuum and the phoneme labels required as responses. However, the scope of such an enterprise would be limited to a narrow branch of speech output systems which are not necessarily the most pertinent for language engineering. Considerable research effort has been expended on going into the details of how the results of assessments like these relate to those employing other psychophysical procedures, which statistical analysis procedures are appropriate, the involvement of memory processes in perceptual decision and so on. Rigour would dictate that all these need to be considered as well as alternative theoretical interpretations of the results of such experiments. Here, as with many of the procedures outlined, followers of one theoretical line stress the importance of different controls in the assessment procedures. Outlining one as a state-of-the-art benchmark is not going to satisfy everyone. The alternative, to present all variants of the procedures and detail their theoretical ramifications, is clearly not possible in a handbook chapter. We will give the general requirements behind constructing experiments as well as representative illustrations of particular types of experiments but do not assume that these represent universal standards.

- 346 Part III: Spoken language system assessment
- 9.1.2 Role of statistical analysis and experimentation in Language Engineering Standards (LES)

In talking about procedural considerations in language engineering, it will help to make things concrete. Let us assume that a client has commissioned development of a speech recognition system (System A) from scratch where expense is no object. It is to be employed in a European country where all inhabitants might want to use it. At the end of the day the client wants to have some idea about how its performance compares with another system on the market (System X). The company is given a free hand when developing the system and would prefer, for convenience purposes, to develop it on the basis of read speech though, as noted, it will eventually have to operate with spontaneous speech. The team assigned to the project decided to develop a system based on Artificial Neural Networks (ANNs).

Some of the questions the team commissioned to do the work may decide to address are:

- 1. How to check whether there are differences between spontaneous and read speech, then make a decision whether the results with read speech apply to spontaneous speech.
- 2. If they find differences between read and spontaneous speech that require them to use the latter, how can they check whether language statistics on a sample of recordings they make to train and test the ANNs is representative of the language as a whole? Whether read or spontaneous speech is used, segments need labelling for training the networks and judges need to be brought in for this purpose.
- 3. What procedures are appropriate for the tasks of labelling and classifying the segments?
- 4. How can the accuracy of segment boundary placement and category classification by the judges be assessed?
- 5. How can improvement during development stages be monitored? This usually involves correct recall of training data by the ANNs. Here segmentation and classification differences between judges (see 2 and 3) might affect assessed recogniser performance. The preceding tests are vital to ensure that the training data is good and that changes in recogniser performance reflect improvements in the architecture, not artefacts of poor training data: An improvement in recogniser performance can be due to a genuine improvement that has been effected or a judge might have made errors and some change allows the system to make the same ones which would then appear to be correct (i.e., the two errors cancel themselves out). Without appropriate assessment of judges' performance, the latter can never be ruled out.
- 6. How are appropriate test data chosen?

These points highlight some of the statistical analysis and experimental procedures that need to feature in language engineering. Moreover, the specific questions raised, though pertaining to a particular issue of concern, are illustrative of many similar problems that language engineers encounter. Now we will set about attempting to provide answers to these (and other) questions.

The remainder of the chapter is organised in five main sections (9.2-9.6). These are (9.2) statistical and (9.3) experimental techniques to ensure that the corpora employed for training and testing are representative, (9.4) assessing speech recognition, (9.5) speaker verification and (9.6) dialogue systems. Sections 9.2 and 9.3 introduce an understanding of statistical analysis and experimentation, and should be read by anyone who does not have background in these subjects. The materials in sections 9.2, 9.3 and 9.4 are specifically focussed on the hypothetical scenario outlined above.¹

9.2 Statistical and experimental procedures for analysing data corpora

In the first part of this section, some fundamental ideas in statistics will be illustrated through selected LES examples. In the subsequent parts, the steps in the scenario that has just been run through will be recapitulated to consider those issues that cannot be addressed until all components introduced are understood. The issues that involve experimentation will then be dealt with.

9.2.1 Statistical analysis

Statistics is the acquisition, handling, presentation and interpretation of numerical data. Language engineers have considerable experience acquiring, handling and presenting data, but less in interpretation. The following will, then, be mainly concerned with the latter.

9.2.2 Populations, samples and other terminology

A *population* is the collection of all objects that are of interest for the task in hand. In the earlier example, all inhabitants of the country are the population. Here everyday use of the term 'population' corresponds with its use in a statistical sense. Though population in a statistical sense can have the same meaning as the geographical sense, it need not be the case. Thus, for instance the population of users of a speaker verification system of a bank would only comprise the clients of the bank. Population does not only refer to humans – for example, the population of /p/ phonemes of a speaker would be all of the instances of that phoneme a speaker ever produces.

¹Though the organisation into these sections is convenient, note that the subdivision into sections is to some extent artificial: The relationship between setting up corpora and testing recognisers is a case of the proverbial chicken and egg – poor performance of a recogniser can be due to training and testing on a poor corpus. In turn, speaker verification and dialogue systems depend to an extent on speech recognition.

A variable ranges over numerical values associated with each unit of the population. Variables are classed as either *independent* or *dependent variables*. An independent variable is one that is controlled or manipulated by the experimenter. So, for example, when setting up a corpus, the experimenter might consider it necessary to ensure that as many females are recorded in the test data as males. Sex would then be an independent variable (independent variables are also referred to as *factors*, particularly in connection with the statistical technique *Analysis of Variance*, *ANOVA* discussed in Section 9.2.6.2). A dependent variable is a variable that the investigator measures to determine the effect of the independent variable. Thus you might need to ascertain whether recognition accuracy (dependent variable) is affected by the sex of the speakers (independent variable).

When a variable is measured on all units of a population, a full census has been taken. If it were always possible to obtain census data, there would be no need for statistics. However, since most language engineering applications (and, indeed, in many other aspects that require measurement), involve very large or infinite populations (such as those illustrated earlier of speakers or phonemes), it is not possible to measure variables on all units: In these circumstances, a finite sample is taken. This sample is used to study the variable of concern in the population. So, if you wanted an idea of the average voice fundamental frequency of men, you might make measurements on a sample of 100 men. This sample is then studied as if it is representative of the population. The statistician is able to provide information about the relationship between variables measured on the sample (here its mean) and, what the investigator is really interested in, the mean voice fundamental frequency of the population.

9.2.3 Sampling

The main problem in treating data statistically is how to ensure the reliability of information about the population obtained from a sample. For this, the sample must be representative. The main requirement to achieve this is to take a *simple random* sample. A sample is a simple random sample if every member of the population has the same chance of being selected as every other member. Thus, if in setting up the ANN recogniser, speakers from the lab are used to obtain training data, the sample would not be simple random: It is unlikely that the members of the lab give all social strata, or females, or people outside working-age, a fair chance of being selected.

9.2.4 Biases

Selection of a sample that is not a simple random sample is one of the main sources of bias in experiments. Bias can be defined as a systematic tendency to misrepresent the population. So, if the ANN recogniser is intended to be used by all members of the population, you cannot select an unbiased sample of speakers from a sample of people recorded just between 9 a.m. and 5 p.m. This would exclude people who are at work; so, if you do this, the result is a biased sample which is not necessarily representative of the target population.

If you take a sample, how sure can you be that if you measure variables such as the mean of the sample is close to the mean of the population? This sort of problem is termed *estimation* and is considered in the following subsections.

9.2.5 Estimating sample means, proportions and variances

Estimation is used for making decisions about populations based on simple random samples. A truly random sample is likely to be representative of the population; this does not mean that a variable measured on a second sample taken will be the same as the first. The skill involved in estimating the value of a variable is to impose conditions which allow an acceptable degree of error in the estimate without being so conservative as to be useless in practice (an extreme case of the latter would be recommending a sample of the same order of magnitude as the population). The necessary background skill is to understand how quantities like sample means, proportions and variances are related to means, proportions and variances in the population. The following notation is used in the discussion: \bar{x} is the sample mean, S is the sample standard deviation, and S^2 is the sample variance; μ is the population mean, σ is the population standard deviation, and σ^2 is the population variance. The abbreviations sd and S.D. are sometimes used for standard deviation; S.E. is used for standard error, Z is used for z-scores, P stands for estimated probability, and p stands for proportion.

9.2.5.1 Estimating means

A fundamental step towards this goal is to relate the sample statistic to a probability distribution: What this means is: if we repeatedly take samples from a population, how do the variables measured on the sample relate to those of the population? To translate this to an empirical example: How sure can you be about how close your sample mean lies to the population mean? Even more concretely, if we obtain the mean of a set of samples, how does the mean of a particular sample relate to the mean of the population. As has already been said, the value of the mean of the first of two samples is unlikely to be exactly the same as the second. However, if repeated samples are taken, the mean value of all the samples will cluster around the population mean; this is usually regarded as an unbiased estimator of the population mean.

The usual way this is shown is to take a known distribution (i.e., where the population mean is known) and then consider what the distribution would be like when samples of a given size are taken. So, if a population of events

has equally likely outcomes and the variable values are 1, 2, 3 and 4, the mean would be 2.5. If all possible combinations are taken (1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, 3 and 4), the mean of the mean values for all pairs is also 2.5 (taking all pairs is a way of ensuring that the sample is simple random). An additional important finding is that if the distribution of sample means (the sampling distribution) are plotted as a histogram, the distribution is no longer rectangular but has a peak at 2.5 (1 and 4, and 2 and 3 both have a mean of 2.5 and none of the means of the other pairs has the same mean). Moreover the distribution is symmetrical about the mean and approximates more to a normal (Gaussian) distribution even though the original distribution gets better. Moreover, this tendency applies to all distributions, not just the rectangular distribution considered. The tendency of large samples to approximate the normal distribution is, in fact, a case of the *Central Limit Theorem*.

This particular result has far-reaching implications when testing between alternative hypotheses (see below). As a rule of thumb sample sizes of 30 or greater are adequate to approximate the normal distribution.

The statistical quantity standard deviation (sd, S.D.) is a measure of how a set of observations scatter about the mean. It is defined numerically as

$$S = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$
(9.1)

Later the related quantity of the variance will be needed. This is simply the sd squared:

$$S^{2} = \frac{(x_{1} - \bar{x})^{2} + (x_{2} - \bar{x})^{2} + \dots + (x_{n} - \bar{x})^{2}}{n - 1} = \frac{\sum_{i=1}^{n} (x_{i} - \bar{x})^{2}}{n - 1}$$
(9.2)

An important aspect of the situation described is that the sample means themselves (rather than the observations) have a *standard deviation* (sd). The sd of the sample means (here the sd of all samples of size two for the rectangular distribution) is related to the sd of the samples in the original distribution by the formula:

$$S.E. = \frac{\sigma x}{\sqrt{n}} \tag{9.3}$$

This quantity is given a particular name to distinguish it from the sd – it is called the *standard error* (*S.E.*). In practice, the standard deviation of the population is often not known. In these circumstances, provided the sample is sufficiently large, the standard deviation of the sample can be used to approximate that of the population and the above formula used to calculate the S.E. The S.E. is used in the computation of another quantity, the *z score* of the sample mean:

$$Z = \frac{\bar{x} - \mu x}{S.E.} \tag{9.4}$$

The importance of this quantity is that the measure can be translated into a probabilistic statement relating the sample and population means. Put another way, from the z score, the probability of a sample mean being so far from the population mean can be computed.

To show how this is used in practice: if a sample of size 200 is taken, what is the probability that the mean is within 1.5 S.E.s of the population mean? Normal distribution tables give the desired area. Here is a section of a table giving the proportion of the area of a normal distribution associated with given values of z (the stippled section in the figure indicates what area is tabulated):



The sketch of the normal distribution is symmetrical and the symmetry is about the mean value (i.e., the peak of the distribution). The z values above the mean are tabulated, and the row with a z value of 1.5 indicates that 0.4332 of the area on the right half of the distribution lie within 1.5 S.E.s above the mean. Since it has already been noted that the distribution is symmetrical, 0.4332 of the area will lie within 1.5 S.E.s below the mean. Thus, the area within 1.5 S.E.s above or below the mean is 0.4332 + 0.4332, or 0.8664. Thus, converted to percentages, approximately 86.6% of all samples of size 200 will have means within 1.5 S.E.s of the population mean. If, as in any real experiment, one sample is taken, we can assign a statement

about how likely that sample is being within the specified distance of the mean.

Another, related, use of S.E.s is in stipulating confidence intervals. If you look at the areas associated with particular z values in the way just described, you should be able to ascertain that the area of a normal distribution enclosed within z values ± 1.96 S.E.s of the mean \bar{x} is 95%. Thus, if the S.E. and mean \bar{x} of a sample are known, you can specify a measurement interval that indicates the degree of confidence (here 95%) that the population mean will be within these bounds. This is between the value 1.96 × the S.E below the sample mean and 1.96 × the S.E. above the sample mean. This particular interval is called the 95% confidence interval. Other levels of confidence can be adopted by obtaining the corresponding z values.

Since this topic is so important, an example is given: Say a random sample of mean voice fundamental of 64 male university students has a mean of 98 Hz and a standard deviation of 32 Hz. What is the 95% confidence interval for mean voice fundamental of the male students at this university? The maximum error of the estimate is approximated (using sample standard deviation S rather than that of the population σ as an approximation, see above) as:

$$1.96 \cdot \frac{S}{\sqrt{n}} = 1.96(\frac{32}{\sqrt{64}}) = 7.84 \tag{9.5}$$

Thus, the 95% confidence interval is from 98 - 7.84 = 90.16 to 98 + 7.84 = 105.84. Often, the confidence intervals are presented graphically along with the means: the mean of the dependent variable is indicated on the y axis with some chosen symbol; a line representing the confidence interval extends from (in this case) 90.16 to 105.84 and it is drawn vertically and passes through the mean.

Before leaving this section, it is necessary to consider what to do when wanting to make corresponding statements about small-sized samples which cannot be approximated with the normal distribution. Here computation of the mean \bar{x} and standard error S.E. proceeds as before. Since the quantity z is used in conjunction with the normal distribution tables, it cannot be used. Instead the analogous quantity t is calculated:

$$t = \frac{\bar{x} - \mu x}{S.E.} \tag{9.6}$$

The distribution of t is dependent on sample size n and so (in essence) the t value has to be referred to different tables for each size of sample. The tables corresponding to the t distribution are usually collapsed into one table and the section of the table used is accessed by a parameter related to the sample

size n (the quantity used for accessing the table is n-1 and is called the degrees of freedom). Clearly, since several different distributions are being tabulated, some condensation of the information relative to the z tables is desirable. For this reason, t values corresponding to particular probabilities are given. Consideration of t tables emphasises one of the advantages of the Central Limit Theorem insofar as one table can be used to address a wide variety of issues rather than is the case for t.

9.2.5.2 Estimating proportions

Here the problem faced is similar to that with means: A sample has been taken and the *proportion* of people meeting some criterion and those not meeting that criterion are observed. The question is with what degree of *confidence* can you assert that the proportions observed reflect those in the population? Once again the solution is directly related to that discussed when estimating how close a sample mean lies to the population mean using z scores. Essentially the z score for means measures:

$$Z = \frac{estimated mean - population mean}{S.E.}$$
(9.7)

The only difference here is that binomial events are being considered (meet/not meet the criterion). Since the mean of a binomial distribution is np (number tested × population proportion) and the *S.E.* is $\sqrt{\frac{pq}{n}}$ where q = 1 - p), the z score associated with a particular sample based on the estimated probability and the population proportion is:

$$Z = \frac{\hat{P} - p}{\sqrt{(pq/n)}} \tag{9.8}$$

Normal distribution tables can again be used to assign a probability associated with this particular outcome.

To illustrate with an example: Suppose that it is expected that as many men will use the ANN system as will women (p (man) = p (woman) 0.5). What size of sample is needed to be 95% certain that the proportion of men and women in the sample differs from that in the population by at most 4%?

$$1.96(=95\%) = \frac{.04}{\sqrt{(.5)(.5)/n}} \tag{9.9}$$

Solving for n gives 600.25. Therefore, a sample of size at least 601 should be used.

Now what are the effects if we want to be more than 4% confident, say if the difference is reduced to 2%. The required sample size jumps to 2401.

9.2.5.3 Estimating variance

The relationship between the variance of a sample and that of the population is distributed as χ^2 (chi squared) with n-1 degrees of freedom.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$
(9.10)

Thus, if we have a sample of size 10 drawn from a normal population with population variance 12, the probability of its variance exceeding 18 is:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$
(9.11)

$$= \frac{9 \cdot 18}{12} = 13.5 \tag{9.12}$$

This has associated with it 9 degrees of freedom. Because χ^2 values are only tabulated for particular probabilities (as with t), the probability can only be estimated for limited probabilities. In this case χ^2 lies between 0.2 and 0.1.

9.2.5.4 Ratio of sample variances

If two independent samples are taken from two normal populations with variance σ_1^2 and σ_2^2 , the ratio of the two variances $(S_1^2 \text{ and } S_2^2)$ has the F distribution:

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \tag{9.13}$$

If the two samples (which can differ in size) from the same normal population are taken, then the ratio of the variances will be approximately 1. Conversely, if the samples are not from the same normal population, the ratio of their variances will not be 1 (the ratio of the variances is termed the *F ratio*). The *F* tables can be used to assign probabilities that the sample variances were or were not from the same normal distribution. The importance of this in the Analysis of Variance (ANOVA) will be seen later.

9.2.6 Hypothesis testing

9.2.6.1 Simple hypothesis testing

Many applications in language engineering require testing of hypotheses. An example from the scenario given in Section 9.1.2 was testing whether there were differences between read and spontaneous speech with respect to selected statistics. If the statistic was mean vowel duration in the two conditions where speech was recorded, we have a situation calling for simple hypothesis testing. This situation is called simple hypothesis testing since it involves a parameter of a single population.

Following the approach adopted so far, the concepts involved in such testing will be illustrated for this selected example. The first step is to make alternative assertions about what the likely outcome of an analysis might be. One assertion is that the analysis might provide no evidence of a difference between the two conditions. This case is referred to as the null hypothesis (conventionally abbreviated as H_0) and asserts here that the mean tone unit duration in the read speech is the same as that in the spontaneous speech. Other assertions might be made about this situation. These are referred to as alternative hypotheses. One alternative hypothesis would be that the tone unit duration of the read speech will be less than that of the spontaneous speech. A second would be the converse, i.e. the tone unit duration of the spontaneous speech will be less than that of the read speech. The decision about which of these alternate hypotheses to propose will depend on factors that lead the language engineer to expect differences in one direction or the other. These instances are referred to as one-tailed (one-directional) hypotheses as each predicts a specific way in which there will be a difference between read and spontaneous speech. If the language engineer wants to test for a difference but has no theoretical or empirical reasons for predicting the direction of the difference, then the hypothesis is said to be two-tailed. Here, large differences between the means of the read and spontaneous speech, no matter which direction they go in, might constitute evidence in favour of the alternative hypothesis.

The distinction between one and two-tailed tests is an important one as it affects what difference between means is needed to assert a significant difference (i.e., support the null hypothesis). In the case of a one-tailed test, smaller differences between means are needed than in the case of two-tailed tests. Basically, this comes down to how the tables are used in the final step of assessing significance (see below). There are no fixed conventions for the format of tables for the different tests, so there is no point in illustrating how to use them. The tables usually contain guidance as to how they should be used to assess significance.

Hypothesis testing involves asserting what level of support can be given in favour of, on the one hand, the null, and, on the other, the alternate hypotheses. Clearly no difference between the means of the read and spontaneous speech would indicate that the null hypothesis is supported for this sample. A big difference between the means would seem to indicate that there is a statistical difference between these samples if the direction in which the means differs is in the same direction as hypothesised for a one-tailed hypothesis or if a two-tailed test has been formulated. The way in which a decision whether a particular level of support (a probability) is

provided is described next.

In the read-spontaneous example that we have been working through, we are interested in testing for a difference between means for two samples where, it is assumed, the samples are from the same speaker. The latter point requires that a related groups test as opposed to an independent groups test is used (see Figure 9.1 on page 358). In this case, the t statistic is computed from:

$$t = \frac{\text{mean of condition}_1 - \text{mean of condition}_2}{S.E. \text{ of differences}}$$
(9.14)

Thus if the read speech for 15 speakers had a mean tone unit duration of 40.2 centiseconds and the spontaneous speech 36.4 centiseconds and the standard deviation of the difference between the means is 2.67, the t value is 1.42. The t value is then used for establishing whether two sample means lying this far apart might have come from the same (null hypothesis) or different (alternate hypothesis) distributions. This is done by consulting tables of the t statistic using n/-1 degrees of freedom (here n refers to the number of pairs of observations).

In assessing a level of support for the alternate hypothesis, decision rules are formulated. Basically this involves stipulating, assuming that the samples are from the same distribution, that if the probability of the means lying this far apart is so low then a more likely alternative is that the samples are drawn from different populations. The "stipulation" is done in terms of discrete probability levels and, conventionally, if there is a less than 5%chance that the samples were from the same distribution, then the hypothesis that the samples were drawn from different distributions is supported (the alternative hypothesis at that level of significance). Conversely, if there is a greater than 5% chance that the samples are from the same distribution, the null hypothesis is supported. In the worked example, with 14 degrees of freedom, a t value of 1.42 does not support the hypothesis that the samples are drawn from different populations, thus the null hypothesis is accepted. It should be noted that support or rejection of these alternative hypotheses is statistical rather than absolute. In 1/20 (5%) cases where no difference is asserted, a difference does occur (referred to as a Type II error, accepting the null hypothesis when it is false) and in cases where a 5% significance level is adopted and differences found, 1 occasion out of 20 will also lead to an error (referred to as a Type I error, rejecting the null hypothesis when it is in fact true).

9.2.6.2 Analysis of Variance

This chapter of the Handbook does not cover all statistical tests that might be encountered, only offer a background and point to relevant material. However, some comments on Analysis of Variance (ANOVA) are called for as it is a technique that has a widespread use in language engineering.

ANOVA is a statistical method for assessing the importance of factors that produce variability in responses or observations. The approach is to control for a factor by specifying different values (or, treatment levels) for it in order to see if there is an effect. It can be thought of as having sampled a potentially different population (different in the sense of having different means). Factors that have an effect change the variation in sample means, where "factor" refers to a controlled independent variable. When the experimenter controls the levels of the factors, this is referred to as *treatment level*.

For example, in the ANOVA approach, two estimates of the variances are obtained: the variance between the sample means, between groups variance, and the variance of each of the scores about their group mean, within groups variance. If the treatment factor has had no effect, then variability between and within groups should both be estimates of the population variance. So, as discussed earlier when the ratio of two sample variances from the same population was considered, if the F ratio of between groups to within groups is taken, the value should be about 1 (in which case, the null hypothesis is supported). Statistical tables of the F distribution can be consulted to ascertain whether the F ratio is large enough to support the hypothesis that the treatment factor has had an effect resulting in larger variance of the between group to the within group means (the alternative hypothesis is supported). Another way of looking at this is that the between groups variance is affected by individual variation of the units tested plus the treatment effect whereas the within groups estimate is only affected by individual variation of the units tested.

ANOVA is a powerful tool which has been developed to examine treatment effects involving several factors. Some examples of its scope are that it can be used with two or more factors. Factors that are associated with independent and related groups can be tested in the same analysis, and so on. When more than one factor is involved in an analysis, the dependence between factors (interactions) comes into play and has major implications for the interpretation of results. A good reference covering many of the situations where ANOVA is used is Winer (1971). Though statistical texts cover how the calculations are performed manually, these days the analyses are almost always done with computer packages. The packages are easy to use if ANOVA terminology is known. Indeed the statistical manuals for these programmes (such as MINITAB, SPSS and BMDP) are important sources which discuss how to choose and conduct an appropriate analysis and should be consulted.



Figure 9.1: Summary of decision structure for establishing what statistical test to use for data

9.2.6.3 Non-parametric tests

Parametric tests cannot be used when discrete, rather than continuous, measures are obtained since the Central Limit Theorem does not approximate

the normal distribution in these instances. The distinction between discrete and continuous measures is the principal factor governing whether a parametric or non-parametric test can be employed. Continuous and discrete measures relate to another taxonomy of scales – interval, nominal and ordinal: interval scales are continuous and the others are discrete. Statisticians consider this taxonomy misleading, but since it is frequently encountered, the nature of data from the different scales is described. Interval data are obtained when the distance between any two numbers on the scale are of known size and is characterised by a constant unit of measurement. This applies to physical measures like duration and frequency measured in Hertz (Hz) which have featured in the examples discussed to now. Nominal scales are obtained when the measures are obtained from symbols to characterise objects (such as sex of the speakers). Ordinal scales give some idea of the relative magnitude of units that are measured but the difference between two numbers does not give any idea of the relative size. The examples discussed below in connection with Likert scales represent this level of measurement. In cases where parametric tests cannot be used, non-parametric (also known as distribution-free) tests have to be employed. The computations involved in these tests are straightforward and covered in any elementary text book (Siegel 1956). A reader who has followed the material presented thus far should find it easy to apply the previous ideas to these tests. To help the reader access the particular test needed in selected circumstances (parametric and non-parametric), a tree diagram for the different decisions it is necessary to make is given in Figure 9.1. Thus, a particular test might require a number of judges to indicate how acceptable they think the synthesiser output is from before to after a change has been made. The dependent variable is the frequency of judges considering whether the synthesiser produced satisfactory output or not before and after the change; for this, a McNemar test is appropriate.

A number of representative questions a language engineer might want to answer were considered at the start of this section. Let us just go back over these and consider which ones we are now equipped to answer. First there was how to check whether there are differences between spontaneous and read speech.

If the measures are parametric (such as would be the case for many acoustic variables), then either an independent or related t test would be appropriate to test for differences. An independent t test is needed when samples of spontaneous speech and read speech are drawn from different speaker sets; a related t test is used when the spontaneous and read samples are both obtained from the same group of speakers.

If the measures are non-parametric (e.g. ratings of clarity for the spontaneous and read speech) then a *Wilcoxon test* would be used when the read and spontaneous versions of the speech are drawn from the same speaker

and a Mann-Whitney U test otherwise.

If you find differences between read and spontaneous speech that require them to use the latter for training data (see application described), how can you check whether language statistics on your sample of recordings is representative of the language as a whole – or, what might or might not be the same thing, how can you be sure that you have sampled sufficient speech? For this, the background information provided to estimate how close sample estimates are to population estimates is appropriate.

The next questions in our list given in the introduction lead on to the second major theme which we want to cover: the general principles behind setting up a well-controlled experiment. The particular experiments that will be considered concern the assessment of the procedures for segmenting and labelling the speech for training and testing the ANNs. The discussion concerning experimental design, etc. will apply to many more situations, however. Once we have an idea what the experimental data would look like, we can consider how to treat the data statistically, which involves hypothesis testing again.

9.3 Experimental procedures

The principles of good experimental design will be illustrated by considering first what procedures are appropriate to provide phonemically-labelled data which is to be used for training and subsequently testing a recognition system. Aspects of these procedures and considerations will apply to a wide variety of experimental situations encountered in connection with language engineering. Finally, it ought to be noted that the performance of a recogniser (to be dealt with in Section 9.4) can only be as good as the data it is trained on, and that assessment results for recognisers will be misleading if the data used for training and testing are in error or if sampling biases are inadvertently introduced by adoption of inappropriate sample selection. Thus, there is a close relationship between the topics discussed here and those in Section 9.4.

9.3.1 Experimental selection of material

The issue of what material to employ applies to speech synthesis as well as recognition. Before concentrating on some issues relevant to recognition, a brief comment is in order about selecting a sample speaker to base a synthesis system on as these have different requirements to those of recognition: In synthesis, basically, you are trying to select one or a small number of speakers whose speech conforms to certain criteria. One reasonable criterion might be that the speaker should be "highly intelligible". Though professional groups like radio announcers or people with voice training would seem like reasonable candidates for this, from a scientific point of view, some metric to check this ought to be applied so that the decision is not subjective. A metric could be developed based on Likert scales (see Section 9.4.3).

If an ANN-based recogniser is to be trained and tested, different requirements are imposed: If the goal is to recognise unrestricted speech by the speakers of a language (as in the example), a statistically representative sample of exemplars is needed. This means that the sample should conform in its statistical properties to whatever population it is supposed to represent. There are two major facets of the representativeness here. An adequate sample of different speakers is needed (see San ection 9.2.5.2 for some estimation methods appropriate for this) and the structure of the materials produced by the selected group of speakers (whether it is spontaneous or specially-constructed material for reading) needs checking to see whether it conforms to unrestricted speech in the language. An explicit formulation needs to be made concerning what would constitute an adequate check on whether the samples are representative: A weak formulation might be to check whether the sample contains all the language's phonemes. A stronger formulation would be to establish whether the sample contains all the language's diphones. An even stronger formulation would be to check whether all the diphones occur with the same frequency as they do in spontaneous samples of speech drawn from the language. The stronger the formulation, the more likely it is that the sample is representative of the language but, on the other hand, the more work is required in obtaining and comparing the sample against the language.

Considering phonemes, first, there are reasons to suppose that checking for these units alone would not be satisfactory. The main problem is that if you only check on whether all phonemes occur, then it is being implicitly assumed that some salient identifying property of a phoneme can be extracted whatever phonemic context it occurs in. This assumption is controversial. Diphones are the observed two phoneme combinations for the language. The advantage of this and related units (such as the demisyllable) is that they include some measure of context over adjacent phonemes. The disadvantage is that there are more of these units than there are phonemes and, consequently, larger samples of speech are needed to ensure representativeness. Ideally, what would be required if these units are to be the basis of a recogniser is a check that all diphones that occur in spontaneous speech occur in the sample. It cannot be assumed that *phonetically balanced* passages control for all diphone contexts. For instance, a short phonetically balanced text does not contain examples of phonemes in all the 900 or so diphone contexts that occur in samples of spontaneous speech. The issue about whether passages should be generated that have, say, diphones with the same frequency as occurs in the language itself is tricky for two reasons: First, the baseline data about diphone frequency is not universally available. Second, there are alternative viewpoints about what would be the best structure for

the frequency of diphones in a sample. On the one hand, it can be argued that the rare diphones are highly informative segments of speech. It might then be advisable to ensure that these occur with the same frequency as the commonest diphones by generating specific passages. The other point of view is that by artificially manipulating the diphone frequencies in this way, the sample is not representative of the language the recogniser would need to work with. This can be illustrated by considering materials that have been used for training Hidden Markov Model recognisers. Sentences in projects such as SPAR were sometimes developed to obtain an instance of each phoneme of English in a small amount of material. However, inspection of some of the sentences shows that they might be difficult to speak (tongue-twisters) and consequently may lead to pronunciation problems (in particular, abnormally timed speech):

- George made the girl measure a good blue vase.
- Be sure to fetch a file and send theirs off to Hove.
- Why are you early you owl?

A speaker is likely to experience difficulty on phonemes in this type of material that he would not encounter when these same phonemes occur in other sentences. To the extent that these sentences behave like tongue twisters, the difficulty encountered would be more acute for certain classes of phonemes (consonants and particularly with plosives) than others (the vowel sounds). A person who still wants to use this material might reply that it is conceivable that these sentences could have been uttered, which is true. However, the discussion of sampling (above) illustrates that they cannot be considered a simple random sample. If it is necessary to use the phonemes in them as instances for training particular phone models, their acoustic properties should be checked statistically against other groups of sentences that also contain these phonemes. This analysis would establish whether there are differences between the acoustic properties in the wider samples and these compressed versions. To our knowledge, these tests have not been conducted. They are essential checks that should be made before a compressed sample is used when the final system is applied to with less restricted materials.

If unrepresentative material is used for training and testing a recogniser, misleading conclusions may well be drawn about its performance. This constitutes a major topic of investigation and arises as follows: Suppose the recogniser is trained on phones marked in passages that are produced atypically due to the phonetic density (such as is the case with the SPAR sentences). When the recogniser fails to recognise a "typical" instance of the phone that differs from the atypical ones that it has been trained on, hasn't it behaved correctly with respect to its training? Conversely, when the recogniser recognises a "typical" instance of the phone that differs from the atypical ones that it has been trained on, hasn't it made an error? What this shows is that if simple random samples are not employed, then, the conclusions about what constitutes both errors and correct outcomes may be misleading.

9.3.2 Segmentation

If speech data are to be used for training phoneme-based recognisers, then data has to be provided in the stored files obtained from human judges which indicate where each phoneme starts and ends (segmentation) and what its identity is (classification). These data are usually obtained from experienced judges. A section of contiguous speech data (possibly augmented by timealigned spectrographic or other materials) is displayed on a graphics screen. The judges have manually-controlled cursors which allow them to select and play sections of the speech.

In order to check the reliability of segmentation and classification judgments, a subset of material has to be checked at least once. This checking can be performed by the same (*intra-judge*) or different (*inter-judge*) judges. Intra-judge reliability will indicate how consistent a particular judge is whilst inter-judge agreement will indicate how consistent different judges are. If different judges are used, it is necessary to ensure that the judges are sufficiently well instructed so that they are making the same type of decision. In experiments, like has to be compared with like. Thus, if an automatic segmentation algorithm is available that works on acoustic patterns, to check the accuracy of the algorithm human judges should be required to do the same thing as the algorithm. Similar considerations apply when comparing manual judgments against those obtained by machine: There is no point, for instance, in the humans locating phonemes based on global spectral properties when the machine is using local audio properties (such as presence or absence of voicing to make these decisions). For assessing recogniser performance, phoneme segmentations are required, not acoustic properties of subsegmental events. Therefore, this dictates what output is appropriate. Note that it is not a foregone conclusion that segmentation will be comparable between humans nor between humans and the algorithms. If, for instance, laryngeal vibration is used as a basis for segmentation, the point where it starts is often not clear cut.

Returning to the procedures that are usually performed when obtaining phoneme labels, once the identity and extent of a phoneme has been ascertained, a permanent record of these parameters is stored in the computer file. Utilities are usually available which allow these parameters to be retrieved and the information about classifications and start and end points of segments to be aligned against the original speech oscillogram, the spectrogram etc. Typically, the judges work through the files from the beginning of the recording to the end.

Written transcriptions of data is frequently performed. These transcriptions have to be aligned against the speech, for example with Dynamic Time Warping (DTW) algorithms. When this is done, it is difficult to establish whether performance problems are due to human errors of judgment or to limitations in the algorithm. In this section, attention is specifically focussed on human judges' performance which excludes obtaining transcriptions and then aligning them.

Another important point about labelling procedures when the labels are entered directly into the computer is that two processes are involved which are, conceivably, logically distinct: This is also implicit in the use of written transcriptions with DTW alignment which will be used to illustrate this point. The DTW alignment locates labels that are given to it in a statistically optimal way (i.e., segmentation). There is no necessary requirement that the transcriber making the written transcriptions has to indicate segment boundaries, only segment categories (i.e., classification). In this case, humans perform one task and machines the other. Other possibilities are to reverse which decision is made by machine and which by the human, or to get the human or machine to do both tasks separately. Though these would be revealing about the quality of the data at segmentation and classification levels and about the influence these would have on recogniser perfomance, they have not been performed to date. The basic precept about conducting experiments applies again: here, the subject should have a clear idea about what decision is being made. In the case of segmenting and labelling when these are conducted together, a situation is encountered where the decisions are mixed (confounded). This makes it difficult during analysis to disambiguate whether, when an error occurs, the error is associated with one decision (say segmentation) or the other (classification).

The procedures of assessing segmentation and classification separately, outlined in the preceding paragraph, are not currently practised. Nevertheless, separate assessment of the segmentation and classification performance of human judges can be made, albeit to a limited extent. The data available to assess accuracy of segmentation are the location of the segment boundaries in time which will involve the bias brought about by confounding decisions. To perform a comparative test between human judges' performance, it is recommended that segmentation performance be obtained from at least two judges as well as the algorithm. Parameters that might be measured are mean difference in boundary location between the two human sets of judgments. The data can then be analysed with ANOVA where the null hypothesis would be that there would be no difference between the human on two occasions (intra-judge) or by two separate judges (inter-judge).

The following factors need to be taken into account when selecting what subsample to perform the segmentation assessment on: whether to choose sections from all speakers in case judges or speakers show specific difficulties, or whether to do complete assessments on selected judges and speakers. Another factor to consider (which can be investigated with the statistical techniques outlined earlier) is what length of sample to take – the sample ought to be at least long enough to contain examples of all phones of interest if phones are going to be used in the recogniser. This will ensure identification of points where speakers have specific difficulties speaking certain phones, or judges have difficulties in locating some phones.

9.3.3 Classification

Most of the detail on background procedures has been given in the section on segmentation. The main issue to be covered is the extent to which "judges" agree about a category to be labelled and what psychophysical effects affect that judgment.

JUDGES

If the judge is a phonetician, he might well be influenced in boundary placement by the sound just heard. For example this sort of expert judge knows the effects of plosives on duration of the following vowel or those of prepausal lengthening. Consequently, this might influence his categorisation of events in a way atypical of the population of listeners from (in our ANN example) the EU country at large. (It is presumed that the recogniser is to be a model of representative listeners, not a model of listeners trained to hear things in ways that might be coloured by alternative theories.)

RANGE EFFECTS

If the judge is to locate the phonemes of a language, some of the judgments will depend upon duration of the events (e.g. /tf/ - /f/ and vowel quality). In addition, different speakers vary their rate and even the same speaker adjusts his rate during utterances. These influences can lead to variable labelling of phonemes based on well-known psychophysical effects which affect human (but not machine) judgments. This makes the machine's task of duplicating human performance a difficult one.

These effects are called *range effects* (Parducci 1965). They are ubiquitous features of human judgment behaviour but here they will be illustrated for the judgment of speech segments in contexts spoken at different rates. Generally speaking, judgments about the attribute value of an event is affected by the range of the attributes in the contextual material presented for judgment at the same time. So, here judgment about the temporal characteristics of an event to determine, for example, whether it is a /tfa/ or /fa/ will be affected by the temporal properties of the rest of the material: a sound will have to be longer to be judged /fa/ in a slow context than in a fast one. Thus, judgments will be influenced by the contextual material. The changes in /tfa/ and /fa/ counts are due to judges being influenced

by the context; they are not due to changes in speaker behaviour. The net effect is a spurious decrease in the $/\int a/$ count when the rate is slow.

9.4 Assessing recognisers

A basic point about assessing recognisers is that you cannot assess how well a recogniser has performed unless it has been trained with data whose reliability are known. Thus, the issues of assessing human segmentation and classification performance have to be addressed.

9.4.1 Baseline performance

When assessing recogniser performance, precautions need to be taken to ensure the sample is simple random: Say a small group of speakers is available and the set includes a speaker who is markedly different from the rest. If the recognisers are trained on a subset of speakers and tested on the remainder, then when the atypical speaker is included in the training set, he will be excluded from the test set. Consequently performance would be reasonably good even though the model trained is not particularly good because it includes the atypical speaker. Conversely, when the atypical speaker is not included in the training set, the model will be good. However, when the model is tested it will produce poorer performance than previously due to the inclusion of the atypical subject. Thus, it is possible to have better baseline performance for a poor model than a good model. It should be clear that simply omitting this subject reduces the chance of the sample being simple random. The only rigorous way round this problem is to ensure that the training and test data contain sufficient numbers of speakers so as to minimise the effects of atypical speakers (Section 9.2).

The basic data for assessing recogniser performance have a similar structure to that obtained during segmentation and classification: For the recogniser, a set of time-varying parameters of the speech are obtained. Recognisers such as those based on ANNs take as input frames which are usually of fixed length and which are comparatively crudely quantised. Such a recogniser produces a classification of a frame or group of frames as output. At the points where the classifications change, a segment boundary has occurred. Thus, there are two basic measures that can be compared with human judgments about what the passage contains – the relative position of segment boundaries, and correspondence between the classifications of the segments. Scoring metrics have been developed which are intended to measure the latter but, once again, they involve problems which are potentially due to implicit segmentation errors.

9.4.2 Progress

The prerequisite for assessing progress is an adequate measure of the errors the recogniser produces and how these reduce over time. Unfortunately, it is not a simple matter to derive a measure of error performance.

As already mentioned, some measures of recognition performance mix up errors of segmentation and classification. Thus, one common kind of list of the types of events that might occur when comparing a human judge's labels with a machine's are:

- CORRECT: Phoneme A occurred at that point according to the transcription and an A was reported during recognition at that point too.
- SUBSTITUTION (MISMATCH): Phoneme A occurred at that point according to the transcription but something other than an A was recognised.
- DELETED: Phoneme A occurred at that point according to the transcription but nothing was reported (not an A nor anything else). This is usually treated as a special subclass of mismatches. However, it could be due to a segmentation rather than a classification error.
- INSERTED: The transcription would lead one, say, to expect two phonemes to occur in some stretch of speech but three (including an A) were recognised. The remaining two phonemes can be aligned with the transcription so it appears that an A was inserted.

But it is not possible to decide whether deleted and inserted phonemes are instances of segmentation or classification errors: a human judge might label a portion of speech as an affricate whereas the machine might indicate a plosive plus fricative. If the machine had used the same segment boundaries as the human, performance might have been equivalent.

The simplest type of error measure is the number of phonemes that the recogniser correctly recognises compared with the number the human judges correctly recognise. A basic (unresolved) problem for this measure is that if humans cannot provide "perfect" classifications, the machine may be receiving noisy data. Specifically in connection with assessing accuracy of classification, for example, the problem is what is the "correct" answer for phones that subjects do not agree on. This raises another issue specifically in connection with a particular technique that has been applied for assessing recognisers. The technique is *signal detection theory* (SDT) and the technique will first be outlined before problems in applying it to assess recognisers (both humans and machines) are discussed.

The basic idea behind SDT is that errors convey information concerning how the system is operating (in this respect, it is an advance on simple error measures). In the signal detection theory model, it is assumed that there is a distribution of activity associated with the event to be detected (e.g. recognition of phoneme A). The recogniser is performing according to some criterion such that if activity is above the criterion, the recogniser (which can be a human or a machine) reports that the phoneme is present
and below the criterion, subjects report that the phoneme did not occur. Usually, the threshold is set so that most but not all activity associated with a signal leads to that phoneme being recognised. Activity associated with the signal distribution above the criterion threshold results in signals being detected (hit) and those below are "missed". This is shown in Figure 9.2.



Figure 9.2: Activity associated with signal distribution

The abscissa is activity level and the distribution (in terms of Standard Normal Deviate units) represents the probability distribution of events associated with the signal (phoneme A) at the various activity levels. The signals associated with other phones are "noise" in relation to the phoneme and they give rise to a distribution of noise activity which influences recognition. The noise distribution represents the probability distribution of activity levels and the criterion activity level is the same as that applied to the signal distribution. Most of the noise distribution on processes associated with good recognisers will be below the criterion but some will be above. When activity associated with the noise distribution below the criterion is encountered, subjects correctly reject this activity as being associated with phoneme A whilst when it is above this criterion, they incorrectly report a signal to have occurred – referred to in signal detection theory as a *false alarm*. This is shown in Figure 9.3.

The criterion is at the same activity level in each case, so the figures combine to give a complete model of the recognition process (see Figure 9.4).



Figure 9.3: Activity associated with noise distribution

The error classes described earlier are associated with the categories needed for a signal detection analysis as follows:

HITS = correct FALSE ALARMS = False + insertions MISSES = Mismatch + deleted CORRECT REJECTIONS = total phonemes - (correct + false + mismatch + deletions)

With the data available in the form of frequency counts of these categories, standard methods can be employed to ascertain (a) the separation between the mean of the noise and signal distributions and (b) the decision criterion that has been applied. These are referred to as d' and β respectively; d' is particularly important in the present context as it is a measure of the discriminability of the signal distribution from the noise distribution which takes into account all the error information available. A work sheet of the calculations of d' and β and tables needed for the calculation are included as Appendix 1.

A way in which the effects of changing the criterion can be seen is by plotting the relationship between hits and false alarms. The trading relationship is referred to as a *Receiver Operating Characteristic (ROC)*.

The problem alluded to earlier in connection with SDT is distinguishing between what is signal and what is noise. In earlier work it has been as-



Figure 9.4: Activity level associated with signal and noise distribution

sumed that human judges are capable of providing the "correct" answers. However, agreement between judges is notoriously low even for gross classifications (for instance, in the stuttering literature, inter-judge agreement about stutterings is as low as 60% for expert judges). The finer level of classification called for here would lead one to expect that agreement about phone classes would also be low.

Possible ways out of this dilemma are (1) improvement in psychophysical procedures and (2) (related) normalisation procedures across judges to obtain some composite level of agreement.

9.4.3 Functional adequacy and user acceptance

Functional adequacy refers to the fact that recognisers have only to perform a limited range of functions (they might not be required to deal with unrestricted speech for example). User acceptance refers to the fact that users might tolerate something that is not perfect. Each of these topics calls for metrics other than percent correct (9.4.2) and can involve subjective judgments on the part of subjects. So, for these topics, it is necessary to consider what is the best way to obtain information from users about the acceptability of a system.

The recommended way of obtaining the information is in the form of a summated rating scale (Likert 1932). These scales are constructed by

Assessment methodologies and experimental design 371

preparing sets of statements designed to measure an individual's attitude about a particular concept (here, for instance, recogniser acceptance). Typically scales are comprised of several different subscales (in assessing user acceptance of a recognition system, these subscales might include response time, format of feedback to user, etc). Respondents indicate the extent to which they agree with each statement by giving a rating (usually between 1 and 5) indicating the extent to which they agree with each statement. In order to counterbalance for response biases, it is usual to phrase questions so that, here for example, questions indicating affirmative user acceptance would lead to low rating responses for some questions and high ratings for others. An example of of question and response format which might be appropriate for assessing user acceptance might be:

1	2	3	4	5
I found the sys- tem easy to use.				Sometimes I ex- perienced difficul- ties in using the system.

These two questions would tend to lead users to use different poles of the rating scale. During analysis, the scale values need to be reversed. The advantages of Likert scales are:

- 1. All questions constituting a concept or subconcept can be summed to give a composite response.
- 2. A quantitative rather than qualitative measure is provided.
- 3. Likert scales are relatively cheap and easy to develop.
- 4. They are usually quick and easy for respondents to complete.

The construction of questionnaires based on Likert's scale format for the items of any identified concept involves going through a sequence of steps:

- 1. Define the concept or set of subconcepts to measure. The literature needs to be reviewed to ratify the concepts identified and check whether others ought to be included. Care should be taken to ensure that these are clearly and precisely defined: A scale cannot be developed until it is clear exactly what is intended to be measured.
- 2. Design the scale.

Scales will be defined for each item. This will be based on Likert's format where appropriate. The format is not appropriate for collecting information on some concepts (principally, demographic details). At this stage response choices will be specified and instructions formulated for the evaluative basis of each item. A pool of items will be generated at this step, which will be subjected to statistical analysis in later steps.

Administration and item analysis.
 Factor analysis may be used for two purposes for validating the scales:

- 372 Part III: Spoken language system assessment
 - 1. Exploratory Factor Analysis. This is used for studying the mulitidimensionality of the Likert scales that underlie a concept. Two aspects of this are (a) to establish the number of factors that best represent the items, and (b) the interpretation of the factors.
 - 2. Confirmatory Analysis: Exploratory factor analysis provides an optimum statistical description of the data. However, scale construction is premised on certain assumptions about what the scale is intended to measure (e.g. response time). Confirmatory Factor Analysis (CFA), may be used to verify the latter hypothesised factor structure. This can be performed using one of the available covariance structure modelling programs such as LISREL (Joreskog and Sorbom 1984) and ESQ (Bentler 1985).
- 3. Validate and produce norms.

The statistical procedures for analysing the factor analyses are available in the standard texts referred to. The outcome of step 3 will typically not produce a sufficiently high level of reliability during this analysis. Step 3 will have to be repeated iteratively until an acceptable level of reliability has been achieved. At that point the questionnaire and normative outputs are available directly from the analysis.

9.4.4 Methodology

9.4.4.1 Application oriented

Assessment as part of the application will lead to speech being encountered that was not involved in setting up the material employed during training and testing.

9.4.4.2 Reference oriented

It has been suggested that perfomance of a newly-developed recogniser be compared against a reference algorithm (e.g. (Chollet and Gagnoulet 1981)). The procedures for comparing performance between the reference and newly-developed algorithms would be similar, and encounter the same problems, as those described in connection with human-human and humanalgorithm performance.

9.4.4.3 Calibrated databases

The procedures for calibrating databases rely in part on checking that the sampling of the corpus is satisfactory (see Section 9.2.3), or involve being able to compare performance against known answers (the problems involved in providing them has been described in Section 9.3).

9.4.4.4 Manipulated, artificial and diagnostic databases

Data may need to be specially constructed in order to test some specific ideas about why performance of the recogniser is poor. This may involve

difficulties in dealing with breathing noises, hesitations, etc. or because difficulty is experienced in recognising particular phonemes or phoneme types. The construction of special data for these purposes needs to bear in mind the concerns discussed above in connection with providing adequate samples of speech (Section 9.3.1).

9.4.5 Experimental design

There has been relatively little work on what psychophysical procedures are appropriate for assessing speech recognisers. Fortunately, there is a second area of investigation which involves similar issues – measuring intelligibility of speech with the hearing impaired – where different psychophysical procedures have been investigated. The procedures, outlined below, require a judgment to be made to stimulus materials. They can be applied directly by language engineers interested in assessing speech output systems where the stimulus materials are synthesised speech.

The situation is more tricky when considering how the procedures might apply to speech recognition (cf. the discussion in Chapter 10): The basic problem is what is the nature of the stimulus material in these cases. One solution is for the recognisers to produce computed output which is then assessed with the procedures. The outputs of the recogniser could be in the form of phoneme labels which would constitute the stimuli. Alternatively, the phonemic output could be converted to speech form using a text to speech system, and the tests could then proceed in exactly the same way as with synthetic speech. However, it needs to be stated that this seems a roundabout procedure. Assume that human speech is read out from text/script. The recogniser also typically outputs text/orthography. Assessment could then consist of straightforward comparison of input/output (in terms of no. of correct words).

In *magnitude estimation*, subjects choose a positive number to represent the subjective magnitude of the intelligibility of the output of different recognisers.

The *rank order procedure*, as its name suggests, would require subjects to place the different recogniser outputs in order of magnitude of increasing intelligibility.

In *paired comparison* as applied to recogniser assessment, subjects judge which of two recogniser outputs has more or less intelligibility.

9.5 Assessing speaker verification and recognition systems

Speaker verification needs to make use of many of the statistical and experimental procedures outlined in previous sections. The principal procedures that are required for assessing speaker verification and recognition systems are those applied to speech recognition processes in the preceding section.

The dependent variable is going to be whether the speaker is verified as a legitimate user of the system or recognised as a particular individual, respectively. The outcome is, then, a discrete variable. Discrete variables are somewhat simpler measures than those considered in connection with acoustic measures of speech recognition (note that it is not claimed that the problems involved are necessarily simpler). The particular types of statistical tests called for will be non-parametric (see Section 9.2.6.3).

As mentioned before, assessing speech recognisers is integrally related to speech corpora; it is for this reason that an example was worked through from scratch addressing both setting up a corpus and assessing recognition. Here only the additional considerations that need dealing with are described; those to do with sampling speakers for recognition and verification and those to do with assessment.

9.5.1 Sampling rare events in speaker verification and recognition systems

Section 9.2.5.2 contains examples about how many speakers are needed to be 95% certain that the proportion of men and women in the sample differs from that in the population by at most 4%. In this case, the events that were being dealt with (men and women) were equally frequent -p and q were the same at 0.5. If p is quite small as it may be in the case of impostors into speaker verification systems (perhaps less than 1% of the users of a system), it would be necessary to obtain very large samples (running in to tens of thousands). In these cases the Poisson distribution has to be used to estimate sample sizes.

9.5.2 Employing expert judgments to augment speaker verification and assessment for forensic aspects of speaker verification and recognition

One important application of speaker verification is in forensic science. The forensic data that might be used can be obtained in a number of ways: Experts might listen to the material and offer a subjective judgment. This judgment is basically whether the speaker is or is not a particular speaker (i.e. discrete measures based on nominal or ordinal scales). There are lots of ways that expert judgments about speakers could be standardised which would enhance the judgments in the ways that, for example, doctors knowledge in formulating diagnosis has been done and incorporated into automatic disease diagnosis expert systems. It is beyond the scope of the current chapter to present this information. Another approach that has been taken to speaker recognition and verification has been through spectrographic measures (a continuous measure, obtained from an interval scale of measurement). To use all these measurements together calls for statistical techniques which can deal with mixes of continuous (parametric) and discrete (non-parametric) measures. Generalised LInear Modelling Techniques (GLIM) allow models to be constructed that involve predicting dependent variables like those required here from mixture of discrete (including binary), and continuous measures (Aitkin et al. 1989). These have not been applied to forensic applications of speaker verification and recognition but seem appropriate for the task. The technique as it would be applied here enables the experimenter to establish which acoustic and subjective factors differentiate one speaker from another. In contrast, Analysis of Variance (ANOVA) can only deal with continuous measures.

9.6 Interactive dialogue systems

Dialogue systems (see also Chapter 13) include aspects of speech recognition and synthesis. Therefore, once again, much of what has been said before is relevant to this topic. Similarly, decisions about trial sizes can be made on the basis of the statistical information provided previously.

Consultations with the interactive systems group raised two main questions in addition to the issues associated with recognition and synthesis: (1) The need for some statistical appraisal of dialogue simulation systems (so-called Wizard of Oz (WOZ) simulations). In particular, how can these provide answers that have been properly tested, and what mechanisms are there for the introduction of these techniques into actual systems? (2) What sort of dialogue metrics are available that could be used for assessing systems?

It is easy to pose these questions but not so easy to provide answers. There really are not clear cut answers for either of these, as researchers involved in these activities have to answer more fundamental questions: with respect to the first question, the research literature has been more concerned with how to set up appropriate methodologies for WOZ rather than using them to deliver definite proposals about a specific system and why. Proposals for specific systems tend to be the subject of internal reports or in conference proceedings rather than in the journals and, so, are hard to get hold of. Research on the second question has been mainly in the realm of qualitative assessments of limited amounts of material analysed manually. Though a handbook on Language Engineering Standards should include advice on the issues enumerated initially, research is not at a stage that it can provide many answers. Rather than say nothing about these issues, we provide a selection of the available information on WOZ and some work which might make dialogue metrics more quantitative.

9.6.1 Wizard of Oz (WOZ)

A description of WOZ has been included in Chapters 4 and 13. Here it is intended to ask what this technique is useful for. The simple answer is setting up simulation of a dialogue system which will allow testing implica-

tions of a required system without committing the investigator to its actual implementation. Itself it is an experimental procedure. The advantage is its flexibility in comparison with a computational implementation. WOZ design brings the *engineer* and *ergonomist* together to realise a complex task. It should be borne in mind that WOZ is a "means to an end" not "an end". Ultimately, if WOZ has done the job expected of it, it will be discarded and a system implemented. The expectations of both engineer and ergonomist need to be clearly specified at the outset.

The engineer expects the WOZ simulation to provide answers as to what the structure of the dialogue system should eventually look like and should therefore be expected to provide certain information necessary to get the ergonomist started. In particular, the ergonomist needs a model for the language used in the dialogue system and on-going advice about any proposed changes. A potential problem at the outset is whether this initial specification of language is seen by the engineer as the core of the work. After initial experimentation the ergonomist will need to go back and request ways in which this and other aspects of the simulation require alteration by the language engineer.

It is necessary to realise that for WOZ to be useful it has to be cheaper than the direct implementation of a system. It should also be a more efficient tool given the comparative flexibility it offers. The engineer should make the ergonomist aware of exactly what time is allowed for development and testing, what call they can have on the engineer's time and expertise, etc. A "searching in the dark" strategy on the part of the ergonomist is an unsatisfactory situation, and an adequate procedure will need to be carefully negotiated.

The aim will be to attempt to come up with some experimental procedures that will allow the engineer and ergonomist to achieve their ultimate goals. The major practical requirement of WOZ is that it provides answers or proposals about what the ultimate system should be like. The major methodological question is somewhat different: How good is the simulation? Statistical and experimental procedures will be focussed on that.

Audio-only simulations are described, followed by some brief comments on current developments of multimodal systems.

9.6.1.1 Audio-only simulations

9.6.1.1.1 Requirements

The language engineer should provide a description of the user dialogue which supports activities for the proposed task. A representation such as a *state transition diagram*, would be appropriate. A second requirement is a performance specification, i.e. a description of exactly what the device is supposed to do.

Once provided with these basics, the ergonomist needs to establish what sort of factors might limit subjects in their ability to work on the assumption that the dialogue is with a machine and, correspondingly, what sorts of factors in the task (the job of the wizard) are likely to facilitate or prevent this pretext.

9.6.1.1.2 Subject variables

Care must be taken to ensure that the observations are made under conditions which will elicit representative performance: the vocabulary, device users and operating environment should be as similar as possible to the device being simulated. This includes making errors similar to those which might occur in the actual device.

A typical set of instructions to the people trying to create the simulation might be as follows:

You are required to transcribe all utterances of the user subject onto the computer using the keyboard. Speak aloud what is displayed by the computer, and your utterances will be transmitted back to a subject. Be careful to read out what appears on the screen, and not just repeat what the subject said. Although your speech will be distorted to make it sound like synthesised speech, try to minimise the inflections in your voice and to speak as consistently as possible, in order to enhance the "mechanical" effect.

9.6.1.1.3 Wizard variables

The requirements associated with the wizard are really those of a good experimental procedure (Section 9.3). The output should be consistent in content, style and pace. Two examples are:

- A given command should trigger the same response from the wizard.
- The response time must comply to the subject's expectations.

Since the job of being a wizard is not easy, wizards may need to be trained to produce predefined replies or menus, etc.

A factors likely to determine variability in wizards is the level of skill exhibited by the system subject, which will include fatigue and individual differences in aptitude. The first factor may be controlled by recruiting wizards who are likely candidates for developing these skills, and by training. Individual differences will be eliminated if only one wizard is employed; however, this advantage may be reduced if the study is large in scope (see section on multimodal simulations).

Since cognitive load is high, two wizard configurations are used in recent studies: one performs the I/O (receives the questions and generates the answers), the second performs "task level processing" (generates the answers to be formulated by I/O wizard). It is considered that the two-wizard setup is more likely to achieve consistency and not increase response time, though these claims need verifying experimentally.

A final important recommendation is that there should be a permanent record of performance. To this end, questions and answers should be tape-recorded.

9.6.1.2 Multimodal

Future interactive systems may require input from more than one modality: Examples would be speech input to generate visual text or voice operated drawing programmes. When WOZ techniques are employed for these applications, the extra factors that need to be considered are:

- TASK COMPLEXITY: The more modalities, the more functions need to be simulated.
- INFORMATION BANDWIDTH: There are many ways of providing input. In addition, the input load may be too high for a single wizard, resulting in his behaviour becoming inconsistent. In such cases, multiple wizard configurations would be essential.
- MULTI-WIZARD CONFIGURATIONS: The need for multi-wizard configurations results in issues about how to organise collaboration between them. Workload must ideally be spread equally. This is difficult since it relies on the subjects' behaviour and, thus, the roles of the wizards may need to change dynamically. For this reason a third (supervisory) wizard may be needed.

Coutaz and associates have spent time developing recommendations for wizard collaboration. They structure a wizard's task in three steps:

- 1. Acquisition (analysis of message).
- 2. Interpretation what the subject's response means in connection with the task faced.
- 3. Formulation (the emission of an answer).

9.6.2 Dialogue metrics

A requirement of dialogue systems is for metrics to measure the successful transaction of the dialogue. One aspect of this concerns turn taking between the participants. Here different ways that dialogue interaction have been measured in conversations between two humans are described.

9.6.2.1 Psycholinguistic metrics

In dyadic dialogue communication between humans, conversation among speakers is characterised by turn-taking: in general, one participant, A, talks, stops; another, B, starts, talks, stops, and so we obtain an A–B–A– B–A–B distribution of talk across two participants. This transition from one speaker to another has been shown to be orderly, with precise timing and with less than 5% overlap.

Sacks and co-workers (Sacks et al. 1974) suggest that the mechanism that governs turn-taking, and accounts for the properties noted, is a set of rules with ordered options which operates on a turn-by-turn basis, and can thus be termed a "local management system". One way of looking at the participant is to see him as a sharing device operating over a scarce resource, namely control of the "floor". Such an allocational system will require minimum units over which it will operate. These units are, in this model, determined by various features of linguistic surface structure: they are syntactic units (sentences, clauses, noun phrases, and so on) identified as turn-units in part by prosodic means.

Other psychologists working on conversation have suggested a different solution to how turn-taking works. According to this other view, turn-taking is regulated primarily by signals, and not by opportunity assignment rules at all. Duncan (1974), for example, describes three basic signals for the turn-taking mechanism:

- Turn-yielding signals by the speaker,
- Attempt-suppressing signals by the speaker, and
- Back-channel signals by the auditor.

These signals are used and responded to in a relatively structured manner. On such a view, the current speaker will signal when he intends to hand over the floor, and other participants may bid by recognised signals for the right to speak.

9.6.2.2 Acoustic-based measures

A disadvantage of dialogue-based metrics is that (like content analysis), they require time-consuming manual analyses. It would be better, therefore, if automatic, acoustic-based procedures could be developed. A potential problem for acoustic-based metrics of dialogue interaction is that often speakers are not acoustically isolated. This need not (and in some available recordings does not) apply over telephone connections and potentially, therefore, for many dialogue interaction systems. These allow acoustic metrics of disruption to be developed which have the advantage that they are automatic. Little work has been done on this topic. Prosodic factors are a major source of turn taking cues, and acoustic metrics associated with these factors (amplitude, pitch and duration) have been measured (cf. Howell 1990).

In this example, Speaker A is interrupted (unsuccessfully) by speaker B. The terms used to describe the various components of an interruption are summarised in Figure 9.5. The ordinate represents activity (which is happening

when the speech is above the baseline).

Telephone systems are particularly useful for this work as they permit acoustical isolation of the two dialogue channels preventing, to an extent, crosstalk. For this reason, interruption patterns like those described above are easily computed.



Figure 9.5: Schematic illustration of terms used in connection with speaker interruption patterns

The types of metrics that can be computed at points of interruption are those associated with prosody – principally amplitude, pitch, segment duration and pauses. Once obtained, these data can be employed to ascertain how these factors are used during dialogue to signal that some response (either from the machine or human) is expected. If the frequency of occurence of particular pitch movements at points of interruption is to be compared across speaking tasks, then a non-parametric statistical test would be needed. If acoustic measures in the vicinity of interruptions are to be compared, a parametric statistical test would be needed.

10 Assessment of recognition systems

10.1 Introduction

This chapter is intended to give the reader information on how to measure the performance of an automatic speech recognition system. In order to do so, the range of types of recognition systems and the applications for the system must first be specified, as this is important for the assessment method.

Much effort in defining standards for assessment of automatic speech recognition (ASR) systems has been invested by the ESPRIT SAM consortium. The reader is therefore referred to the final reports of SAM (SAM 1992).

10.1.1 Classification of recognition systems

A number of parameters define the capability of a speech recognition system. In Table 10.1 these parameters are categorised. The classification made here is based upon the typical design considerations of a recognition system, which may be closely related to a specific application or task. In general, these parameters are one way or another fixed into the system. For each of the categories, the extremes of an easy and difficult task, from the recogniser's point of view, are given.

- VOCABULARY SIZE The vocabulary size is of importance to the recogniser and its performance. The vocabulary is defined to be the set of words that the recogniser can select from, i.e. the words it can refer to. In cases where there are few choices the recognition is obviously easier than if the vocabulary is large. The adjectives "small", "medium" and "large" are applied to vocabulary sizes of the order of 100, 1000 and (over) 5000 words, respectively. A typical small vocabulary recogniser can recognise only ten digits, a typical large vocabulary recognition system 20 000 words.
- SPEECH TYPE There is a distinction between "isolated words", "connected words" and "continuous speech". For isolated words, the beginning and the end

Parameter	Easy task	Difficult task
Vocabulary size	small	$\operatorname{unlimited}$
Speech type	isolated words	continuous speech
Speaker dependency	speaker dependent	speaker independent
Grammar	strict syntax	natural language
Training method	multiple training	embedded training

Table 10.1: Classification of speech recognition systems

of each word can be detected directly from the energy of the signal. This makes the job of word boundary detection (segmentation) and often that of recognition a lot easier than if the words are connected or even continuous, as is the case for natural connected discourse. The difference in classification between "connected words" and "continuous speech" is somewhat technical. A connected word recogniser uses words as recognition units, which can be trained in an isolated word mode. Continuous speech is generally connected to large vocabulary recognisers that use subword units such as *phones* as recognition units, and can be trained with continuous speech.

SPEAKER DEPENDENCY The recognition task can be either speaker dependent, or speaker independent. Speaker independent recognition is more difficult, because the internal representation of the speech must somehow be global enough to cover all types of voices and all possible ways of pronouncing words, and yet specific enough to discriminate between the various words of the vocabulary.

For a speaker dependent system the training is usually carried out by the user, but for applications such as large vocabulary dictation systems this is too time consuming for an individual user. In such cases an intermediate technique known as *speaker adaptation* is used. Here, the system is bootstrapped with speaker-independent models, and then gradually adapts to the specific aspects of the user.

GRAMMAR In order to reduce the effective number of words to select from, recognition systems are often equipped with some knowledge of the language. This may vary from very strict syntax rules, in which the words that may follow one another are defined by certain rules, to probabilistic language models, in which the probability of the output sentence is taken into consideration, based on statistical knowledge of the language. An objective measure of the "freedom" of the grammar is the *perplexity*, which measures the average branching factor of the grammar. The higher the perplexity, the more words to choose from at each instant, and hence the more difficult the task. See Chapter 7 for a detailed discussion on language modelling. An example of a very simple grammar is the following sentence-generating

	shall		homo	
Ι	want to	go	to school	(10.1)
	should			

which can generate only six different sentences, which vary in the number of words.

For an example of statistical knowledge, consider the word *million* being recognised. If the *domain* is financial jargon, one can make a prediction of the next word, based on the following excerpt of conditional probabilities:

million acres0.00139million boxes0.00023million canadian0.00846million dollar0.0935

syntax:

	$\operatorname{Parameter}$	easy task	difficult task
Pre:	Vocabulary choice Talking style Recording conditions	distinct words read speech constant energy level undisturbed speech	similar words spontaneous speech fluctuating level deteriorated speech (e.g. stressed,
Post:	Electrical character- istics	wide bandwidth good transmission no noise	small bandwidth unreliable channel quality noise

Table 10.2: Conditions of speech

million dollars 0.642 million left 0.0000081

There are almost two out of three chances that the word following *million* will be *dollars* (at least, within the domain of the Wall Street Journal (WSJ). These numbers were calculated from 37 million words of texts of a financial newspaper (the WSJ).

TRAINING The way an automatic speech recognition system is trained can vary. If each word of the vocabulary is trained many times, the system has an opportunity to build robust models of the words, and hence a good performance should be expected. Some systems can be trained with only one example of each word, or even none (if the models are pre-built). The number of times each word is trained is called the number of *training passes*.

Another training issue that defines the capability of a system is whether or not it can deal with embedded training. In embedded training the systems is trained with strings of words (utterances) of which the starting and ending points are not specified explicitly. A typical example is a large vocabulary continuous speech recognition system that is trained with whole sentences, of which only the orthographic transcriptions are available.

10.1.2 Speech quality and conditions

The conditions under which a recogniser is usedgreatly influence on its performance. Speech quality can be characterised by various properties. There is a distinction between pre-production factors, which influence the way speech is produced, and post-production factors, which influence the way the speech is transmitted from the mouth of the speaker to the recognition system. We have summarised some of the conditions in Table 10.2.

- 384 Part III: Spoken language system assessment
- VOCABULARY CHOICE Within the vocabulary, words can be chosen to be acoustically very distinct, or very similar. One would choose the former for an application (e.g. a set of control words), while for diagnostic purposes the latter serves very well (e.g. CVC-words, see Section 10.3.4).
- TALKING STYLE Firstly, a distinction is made between read speech and spontaneous speech. The former is somewhat unnatural, as there are only few circumstances in which speech approaches this quality, but it has been used in evaluation of speech recognition systems for a long time because it is relatively easy to define and reproduce. Spontaneous speech comes in a variety of flavours, but it generally consists of a much less well-defined grammar, and contains errors, corrections, mispronunciations, and stronger prosody. Secondly, the level of the speech can vary. When the level varies strongly within a short time frame (e.g. the distance between microphone and mouth may not be constant) this is called a *large dynamic range*. On a more global scale, the speech itself can be influenced by the speech level, i.e. the speech can range from "whispering" to "shouting".
- RECORDING CONDITIONS The recording conditions may vary. One of the most important quantities in this respect is the signal-to-noise ratio (SNR). Databases are often recorded "clean" (high SNR), and adverse conditions, such as environmental noise and crosstalk are added to the signal in a later stage. However, for some conditions such an approach is not valid (e.g. with the Lombard effect), and the recordings have to be made under realistic conditions.
- ELECTRICAL CHARACTERISTICS The bandwidth is of some importance to the recognition performance. In principle, limited bandwidth contains less information about the speech, and can hence make the recognition task more difficult. However, some recognition systems may limit the bandwidth to telephone speech on purpose even if wide band speech is available because band limiting has the advantage of reducing the amount of data while keeping most of the speech information. In this way, some trivial filtering of noise outside the typical speech spectrum is obtained. Another "electrical characteristic" is the transmission channel quality. Obviously, non-ideal transformations of the signal, such as non-linearities, ticks, echoes, reverberations and drop-outs, will have a degrading influence on the recognition performance.

10.1.3 Capability profile versus requirement profile

Some ways to classify a speech recognition system and speech quality have been described. Generally, a recognition system can only cope with some of the parameters in the "difficult" region, and must yield towards the other conditions. For instance, a large vocabulary speaker independent continuous speech recogniser only performs well on undisturbed, read speech of a known domain (vocabulary and grammar), and a speaker-independent telephone speech real-time recognition system may be designed to recognise digits only. In this way, the capabilities of a particular recogniser with respect to all the conditions define a *capability profile*.

It is a very good idea to realise that the requirements of a particular application define a similar *application requirement profile* which is generally different from the capability profile of most recognition systems.

Recommendation 1

Before you choose your set of recognisers to assess, define the requirements of the application, if any, and compare those with the specifications of the recognition systems.

10.1.4 Assessment purpose versus methodology

The purpose for which the recogniser can be assessed covers a wide range. It varies from "application oriented", (how well is this system suited to a particular use, or which system is the best for my application) to "focussed on development" which is a diagnostic approach (what are difficult phonemes to distinguish, what are ideal parameter settings). For each purpose one or more methodologies exist.

In the SAM final report (SAM 1992) a range of assessment methods are described:

- representative databases
- reference systems
- specific calibrated test databases
- diagnostic methods with a specific vocabulary
- artificial test signals

The methodologies have been ordered from application oriented to diagnostic, thus suggesting a natural mapping towards the purpose of the assessment. The first method, using representative databases, is an expensive field experiment and difficult to reproduce, but gives the best measure for real performance. The last method, using artificial test signals, is less expensive and reproducible, but gives no absolute measure of performance. In Section 10.3 we will treat the various assessment methods in greater detail.

10.2 Definitions and nomenclature

10.2.1 The performance measure as percentage

In speech recognition assessment, score measures such as accuracy, error rate, substitution rate, etc. are commonly expressed as percentages. This tends to lead to various factors of 100 in the formulas, which sometimes makes them less transparent. In order to define simpler formulas, the percent is to be interpreted literally, i.e. "% = 1/100." Thus, one can view the percent as a dimensionless unit, of value 0.01. To convert a fraction to a percentage, multiply by 100 (i.e. divide by the "unit" %).

10.2.2 Recognition score

Basically, the assessment of an automatic speech recognition system is very simple: you take some speech material, train the system if that is required, have the recognition system recognise the speech, and compare the results to a written transcription of the utterances. The way this is carried out, depends on the particular system, and the purpose of the assessment (see sections 10.5 and 10.6 for two typical ways to do this). For instance, consider the assessment of a phone-based recognition system. If the purpose is diagnostic, concentrating on the acoustical part of the system, the scoring algorithm should be based upon phoneme alignment. For benchmarking purposes, however, a simple word alignment is preferred, and scoring can be based upon word error rates and sentence error rates.

10.2.2.1 Isolated word scoring

The definition of the *error rate* E of a system is not so simple. In words, the error rate is defined as "the average fraction of items incorrectly recognised". Here, an item can be a word, a subword unit (e.g. a phone), or an entire utterance. An average is a statistical property, so experimentally we can only measure an *estimator* for the property, based on observation of a specific sample. The definition of the estimator is simplest for an isolated word recognition system:

$$E = \frac{N_E}{N}.$$
(10.2)

Here N is the number of words in the test sample and N_E the number of words incorrectly recognised. The latter can be further subdivided into the contributions:

$$N_E = N_S + N_D. \tag{10.3}$$

Here, the subscripts S and D are the number of words substituted and the number of words incorrectly rejected (deletions). For these classes of errors the fractions can be defined separately,

$$S = \frac{N_S}{N}, \qquad D = \frac{N_S}{N}. \tag{10.4}$$

It is customary for isolated word recognition systems to express the error rate in its complementary quantity, the fraction of correct words C = 1 - E. It is the fraction of words correctly recognised, and its estimate is

$$C = \frac{N_C}{N} = \frac{N - N_S - N_D}{N}.$$
 (10.5)

This measure does not include so-called *insertions* (see the next section), which are only defined for connected word recognition.

For isolated word recognition systems, there is another measure besides the fraction of correct words, which is also of importance. It describes the capability of rejecting an input word that is not in the vocabulary and the sensitivity to non-speech events. If a recogniser outputs a word when there is no specific input, this is called a *false alarm*. In conditions where there is no speech input, the number of false alarms will most likely scale with time, and the correct measure would be the *false alarm rate* f,

$$f = \frac{N_F}{T} \tag{10.6}$$

expressed in events per second. Here N_F is the number of false alarms observed in a time T. Under the condition that there are many input words not in the vocabulary (as is the case in word spotting systems) the number of false alarms is most likely to scale with the number of input words, and hence an estimator for the *false alarm fraction* F is

$$F = \frac{N_F}{N_{\text{oov}}},\tag{10.7}$$

where N_{oov} is the number of out-of-vocabulary words.

As a last measure, there is the *response time* t_r . It can be defined as the average time it takes to output the recognised word after the input word has been uttered.

In conclucion, the isolated word recogniser has four different performance measures, S, D, f, F and t_r . One can try to combine these measures into one "figure of merit", but the weights to the different quantities depend on the application. The combination of substitutions and deletions are often combined to the error rate E.

10.2.2.2 Connected or continuous word scoring

For a connected word or continuous recognition system the measures of performance are more complicated. Because the output words are generally not time-synchronous with the input, the output stream has to be *aligned* with the reference transcription. This implies that classifications such as substitutions, deletions, words correct and false alarms can no longer be identified with complete certainty.

For these reasons, the term "false alarm" is replaced by the term "inserted word" or "insertion", with the corresponding symbol I and the estimator

$$I = \frac{N_I}{N},\tag{10.8}$$

where N_I is the number of insertions according to the alignment procedure. Because the absolute identification of errors is lost in the alignment procedure, the insertions are generally included in the error rate, so that for connected word recognition, performance is expressed in the total word error rate

$$E = S + I + D = 1 - C - I.$$
(10.9)

Note that this error measure can become larger than 1 in cases of extremely bad recognition.

Often, one defines the *accuracy* of a system

$$A = 1 - E = 1 - S - I - D. (10.10)$$

Note that this is *not* just the fraction C of words correctly recognised, because the latter does not include insertions.

The actual measurement of the quantities through alignment is difficult. See Chapter 10.6 and Hunt (1990) for a discussion about alignment. In the above formulas the three types of errors (S, I, D) have equal weight. Depending on the application, one can assign different weights to the various kinds of errors. Hunt (1990) introduces the concept "figure of merit" for connected word recognisers and discusses the effects of different weights on the alignment procedure.

10.2.3 Confusions

For isolated word recognisers, we can define a more specific measure than the various contributions to the error rate alone. The class of substitutions can be divided into all possible confusions between words. The confusion C_{ij} is defined as the probability that word *i* is recognised as word *j*. (Incidentally, the value C_{ii} is the fraction of times word *i* is *correctly* recognised.) These probabilities can be estimated from a large test sample in the same manner the basic error rates are measured, by measuring the number of times the confusion took place:

$$C_{ij} = \frac{N_{ij}}{\sum_{i'} N_{i'j}},$$
(10.11)

where N_{ij} is the number of times word j was recognised on the input word i. The confusion matrix gives more detailed information than the error rates, but has much worse statistics, as the numbers involved are normally low. If we want to include insertions and deletions in this matrix, a null word i = 0 should be added (formally not in the vocabulary), so that the row C_{0j} contains false alarms, the column C_{i0} the deletions, and $C_{00} = 0$. From this expanded confusion matrix, the error rate can be calculated from the diagonal of the matrix, i.e. $E = 1 - \sum_i C_{ii} = \sum_{i \neq j} C_{ij}$. The elements C_{ij} for $i \neq j$ are called the off-diagonal elements.

10.2.4 Vocabulary

Generally, an automatic speech recognition system can recognise only a limited number of words (or more generally items). The set of all words

the recognition system has been set up to be able to recognise is called the *vocabulary* V. Similarly, one can define the test vocabulary V_t as the set of words appearing in the evaluation test. A word w is called *out-ofvocabulary* (OOV) if it is present in the test vocabulary but not present in the recogniser's vocabulary, i.e. if $w \in (V_t \setminus V)$.

10.2.5 Analysis of Variance design

In an analysis of variance (ANOVA), cf. Chapter 9, the contribution of various parameters on the variance in scoring result is determined, including the interaction of parameters. This is carried out by dividing the values of the parameters in ranges or levels. Thus each parameter p has levels $l_1^p \dots l_{N_p}^p$. An *M*-dimensional space is defined by the Cartesian product of the *M* parameters,

$$p_1 \otimes p_2 \otimes \cdots \otimes p_M. \tag{10.12}$$

This space is divided into cells $c(l^1, l^2, \ldots, l^M)$, the number of cells being

$$\prod_{p=1}^{M} N_p. \tag{10.13}$$

10.3 Description of methodologies

10.3.1 Representative databases

A representative database consists of data collected "in the field." It is a collection of speech material that is very specific to the application: the same speakers in representative recording conditions, and similar words for all conditions the application is being used for. This is the ultimate test for evaluating the recognition system to the very specific conditions and requirements of the application.

The procedure may lead to a decision whether or not to actually use the system. However, often this is not a good starting point for assessment. This is because many recognition systems have to be trained or parameters have to be tuned so as to optimise recognition performance. Disadvantages are that during field experiments many parameters are uncontrolled, or simply things go wrong and cannot be repeated. Also, recording a representative database is very time consuming and expensive. One of the most important principles of recogniser assessment is that one *cannot use a test database more than once*, for a particular system, because as soon as there is feedback from the assessment results to the training state of the system, the system is actually trained to perform well for that specific test, is because the assessment test is often only a small sample of the material that the system is actually going to be used for in the application. This

sample can only be representative of the application if it "has never been seen".

Recommendation 2

Be somewhat reserved towards directly using all the available representative test material for evaluation assessment. You may want to run some other tests before you do this.

The recognition score gives you an idea on how well the system will perform for the selected application. It is important that the test utterances are a representative sample of the application. Conditions to which the sample should conform, are:

- The vocabulary should be representative of the application; frequency of words should occur similarly.
- Other language technological parameters, such as the fraction of out-ofvocabulary words, perplexity of the sentences/sentence grammar, should match the use in the application.
- The speakers should be representative, i.e. the gender ratio, age, speaking rate, dialect etc. should reflect the end users of the application.

Most of these points apply only to the assessment of large vocabulary systems.

10.3.2 Reference methods

The previous methodology will give a measure of performance which is typically the accuracy, word error rate or sentence error rate. This is but one measure of the performance. Another measure can be based upon making the task more difficult and seeing how much difficulty has to be added in order to obtain a certain level of word error rate. (One of the problems with the final application assessment can be that the word error rate is so low that the test must be made very large in order to get results that are statistically significant.) Vice versa, the reference system (i.e. the system to which the recognition system is compared) can get a more difficult task. An example of the latter is the measure introduced by Moore (1977), the "Human Equivalent Noise Ratio", HENR. The performance of a system is compared to human performance scores. For humans, noise is added in order to decrease the scores. The signal-to-noise ratio at which the human score is equal to the recogniser's score, is defined to be the HENR of the recognition system. The advantage of this method is that the performance measure is relatively independent of the test vocabulary, and that it gives a comparison to human recognition by definition. A disadvantage is that the method is very laborious, as the human calibration has to be carried out for each new test database, with various subjects.

Also, a reference speech recognition system can be used as a benchmark. Chollet and Gagnoulet (1981) developed such a software-based recogniser for assessment purposes.

10.3.3 Specific calibrated databases

A more diagnostic approach to compiling a test database has been proposed by Peckham and Thomas (1990). Here, an Analysis of Variance method (ANOVA) is used in order to measure the influence of specific parameters of the input signal to the recognition score.

A number of parameters that can influence the recognition score are defined; these are typically speech parameters such as speaking rate, fundamental frequency, language parameters such as perplexity, sentence length, and quality parameters such as signal-to-noise ratio. For each of the parameters whose influence is to be measured, a number of "levels" is defined, e.g. "high" and "low" (perplexity) or "0–10", "10–25" and "25–50" (dB signal-to-noise ratio). The database then should cover all cells of the design with a sufficient number of observations.

Because the number of parameters M readily increases the total number of cells, M is practically limited. Some of the parameters may be uncontrolled (speaking rate), others may be controlled (signal-to-noise ratio). For uncontrolled parameters it can turn out that not all cells are filled with enough observations; in that case one may have to reduce the number of parameters in the analysis.

When all of the parameters in the database have been categorised into groups, the database is called calibrated.

10.3.4 Diagnostic methods with a specific vocabulary

If the acoustic part of a recogniser is to be assessed, one can use specific vocabularies that concentrate on certain linguistic features. One method was proposed by Steeneken (1987), where the test vocabulary consists of so-called CVC-words (consonant-vowel-consonant). Within the test vocabulary, words differ only in the first consonant (CVC), in the vowel (CVC) or in the final consonant (CVC). Thus, by measuring the confusion matrix, one can get diagnostic information on what consonants (or vowels) are hard to distinguish. This CVC-database assessment method is based upon speech intelligibility evaluation.

One of the major advantages of using a CVC-type database is that the recognition scores will be generally low. This might be counterintuitive, but for diagnostic and development purposes this is useful, because with a relatively small test the recognition score measurement gets reasonably accurate. For instance, if the purpose of the recognition test is to tune some technical recogniser parameters (such as energy threshold), one wants to get a reasonable recognition score that does not saturate to 100 %. In this way, the

change of a parameter will be apparent on a small test set.

Another advantage of the use of CVC-databases is the small size of the vocabulary. For Dutch CVCs, for instance, lists of 17 CVC, 15 CVC and 11 CVC are representative for the language. Such a small vocabulary allows to study confusions accurately, but also makes a quick tune-test cycle possible.

As a result of the SAM effort, a speech database on CD-ROM called EUROM-1 has been produced, which contains CVC-tests for various languages. There are also embedded CVC-tests, where the test words are embedded in carrier sentences (Steeneken 1991). This can be used to test both word spotting systems and connected word recognisers.

Another approach was made by Simpson and Ruth (1987a,b). Their test set is based on Phonetic Discrimination with 100 words (PD-100). The test words have been designed to have minimum difference in phonetic respects for pairwise comparison. Related to the Diagnostic Rhyme Test (DRT), the response is closed (i.e. there are forced choices) and this may lead to misleading results.

10.3.5 Artificial test signals

Although a separate class of assessment method has been defined for artificial test signals, no methods are known that are based upon this. One could conceive of speech synthesis systems that can reproducibly generate speech of desired quality and characteristics.

10.4 Parameters

We have already described various classes of parameters that are relevant to the assessment of automatic speech recognition systems. In this section, the following classes will be addressed:

- Pre-production parameters
- Post-production parameters
- Linguistic parameters
- Recogniser-specific parameters
- Assessment parameters

10.4.1 Pre-production parameters

These parameters include everything that influences the speech before it is uttered. They are typically speaker-specific and include speaking style, Lombard effects, dialects, mental and physical health, speech impairment, etc.

PITCH OR FUNDAMENTAL FREQUENCY The main variability in this parameter is between male and female speakers, the pitch of the women being roughly one octave higher than that of men. Typical values are 200 Hz for female speech and 100 Hz for male speech (Thielen 1992). The pitch of children is so different that they are often treated as "the third sex".

- SPEAKING RATE This parameter can be measured in words/minute for representative words, or as a ratio with respect to the average of all speakers.
- LONG TERM SPEECH SPECTRUM This is actually a set of parameters; a representation can be the average energy in third octave bands.
- TALKING STYLE This includes amplitude (from whisper to shouting), fluency, clarity, etc. These are generally no objective measures.
- $\label{eq:Miscellaneous} \mbox{Miscellaneous Other characteristics, such as dialect, physical health, speech impairment.}$

10.4.2 Post-production parameters

This set of parameters can still be changed after the database has been recorded. This can often be done in only one way, e.g. the signal-to-noise ratio can only be decreased.

- Microphone characteristics. These are output voltage/input pressure, bandwidth, linearity, impedance, directional sensitivity. The microphone type (manufacturer, model number) and speaking distance can be important parameters.
- Acoustic environment of the reader (reverberation, echoes, noise)
- Signal-to-noise ratio (dB)
- Signal level (voltage RMS or dB V)
- Transmission channel characteristics. This includes bandwidth, clipping, bit-errors, etc.

10.4.3 Linguistic parameters

This set of parameters describe the words appearing in the test sentences.

- Vocabulary size (number of words)
- Fraction of out-of-vocabulary words
- Sentence perplexity (see Chapter 7)
- Phonetic similarity between words in the vocabulary

10.4.4 Recogniser specific parameters

Often speech recognition has some tunable parameters that influence performance. It is impossible to give an exhaustive list of possible recogniser-specific parameters. Typical parameters are: energy thresholds for word detection, minimum time gap between words, acceptance threshold, noise rejection and input gain. Sometimes the character of the recogniser can be tuned, by setting parameters such as the number of states in the Hidden Markov Models that are used to build the word models.

RECOMMENDATION 3

It is important to make a list of all parameters and record the values of these before a training or testing session.

It may be difficult to find starting values for these recogniser-specific parameters; in this case the system's manual or the company's help-desk should provide this information.

10.4.5 Assessment parameters

The following list contains parameters that can be changed according to the type of test one uses. These parameters control the level of diagnostics, representativeness, the accuracy of the results, etc.

- SPEECH MATERIAL This is in fact the database choice, as described in Section 10.3. There is a long list of possible speech material to be used. Some frequently used databases are:
 - ISOLATED WORDS for an isolated word recogniser (digits, numbers, the (spelling) alphabet, task related words, etc.)
 - CONNECTED WORDS OR SMALL SENTENCES for a connected word recogniser (digit strings, connected task words according to a syntax, etc.) ISOLATED CVCS for a diagnostic test of an isolated word recogniser
 - CVCS IN CARRIER SENTENCES for diagnostic tests of a connected word recogniser
 - NEWSPAPER SENTENCES for benchmark evaluation of continuous speech recognisers
- SPEAKER CHARACTERISATION Each speaker has some associated properties, such as sex, age, dialect, profession, etc. Some control over these properties can be obtained by selecting the test speakers or specific material in the database.
- NUMBER OF SPEAKERS This parameter is relevant for speaker-independent systems. The variability of speech recognition scores is known to be very dependent on the speaker. Apparently, speakers can be classified as "goats" (low recognition scores) and "sheep" (high recognition scores). Because knowledge of this classification is often not available a priori, many speakers are necessary for a benchmarking evaluation. For a speaker independent recognition system, 20 speakers is considered to be a reasonable amount, but this depends very much on the variance within the individual speaker scores. A sufficient number of speakers allows estimation of the variance in score due to the speaker variability, and significance can be tested using Student's t-test.
- TRAINING METHOD The training method is determined by the possible application. Some applications (typically ones with a large vocabulary) might demand that the complete vocabulary is trained only once or twice for each user. For a dictation system (with "unlimited" vocabulary) one may have to use a pre-trained system. Other applications (e.g. command-and-control)

might assume more effort from the user. If the assessment is application oriented, a representative training should be used.

A relevant parameter for isolated or connected word recognisers is the number of training sessions. Prediction of the performance as a function of number of training sessions may optimise the use.

Recommendation 4

For determining the minimum number of training sessions, carry out a small scale experiment, and make an estimate of the variance in the scores.

For a large vocabulary continuous speech recogniser, the training effort is characterised by the total training time.

- GRAMMAR Often recognition systems are equipped with some kind of a grammar that specifies what the word order of recognition can be. Examples of these are:
 - WORD PAIR In a word-pair grammar (regular grammar), for each word in the vocabulary, a list of words that can possibly follow that word is given. This information can be specified in V^2 bits, where V is the vocabulary size.
 - SYNTAX WITH NODES In a syntax with nodes (context-free grammar), words in the vocabulary are divided into groups. Each group is characterised by a *node*. The syntax defines what nodes may follow other nodes, not specifying which word within each node actually fits the input.
 - *n*-GRAM In an *n*-gram grammar, statistics on the probability of occurrence are given. For n = 1 we speak about a *unigram* grammar, and then for each word of the vocabulary the relative frequency of occurrence is given. Put simply, when the recogniser is in doubt between two possible words, it can use this information to choose the most frequently occurring one. This concept is expanded to bigrams, where the probability that two words occur successively is defined. This concept can be expanded for sequences of n words (see Chapter 7). The word sequences with highest probability, including both the acoustic match and the *n*-gram probabilities, is chosen by the recogniser.

If the automatic speech recognition system is tested with grammar, the input speech should actually match the grammar. For a strict grammar, such as a word-pair grammar or a syntax with nodes, no sentences that are not in accordance with the grammar should be used in assessment if the purpose is benchmarking. However, it is of interest to study the recognition output for ungrammatical speech input, which tests the rejection capability. For probabilistic grammars, the perplexity of the test sentences should match that of the "test set" that was used to generate the grammar, if the purpose of assessment is benchmarking.



Figure 10.1: Simplest set-up for interactive testing of a word recognition system

10.5 Experimental design of small vocabulary word recognition

Although there are many different capability profiles of automatic speech recognition systems, we will now concentrate on two examples that span a wide range of systems, namely the isolated word recogniser and the large vocabulary continuous speech recogniser in the following section.

We will implicitly assume that the isolated word recogniser has some other feature that makes it attractive for an application: "real-time" processing of the speech input. This does not necessarily mean that the recognition process happens in 1 : 1 real time, but we merely want to exploit the fact that isolated word recognition systems can give one word of output on each input word, so that the assessment process can be interactive.

10.5.1 Technical set-up

A typical set-up for testing a real-time word recogniser, is shown in Figure 10.1. The set-up consists of a playback device, the recognition system and a controlling device. Note that for the playback and controlling device, a single person (the experimenter) would be sufficient, but it is wise to try this to get a feel for the recognition process.

Recommendation 5

Before you try to automise the set-up, experiment a little with the system to get a feeling for how it works.

Depending on the level of automation, you can choose for the playback device:

- a microphone and amplifier,
- an analog recording device,
- a digital recording device,
- a computer with mass storage and a sound interface (digital-to-analog converter).

In most cases, the last option is chosen, because of the reproducibility and potential for automation. It also allows for the control of the time of silence between words, added noise, etc. If the recognition system has digital input, the analog path can be avoided completely. The functions of controlling and speech generation can easily be performed by the same computer.

One has to take care, however, that for a connected word recognition system, no "bursts" of speech should be generated, because a connected word systems is continuously "listening". Performance will be influenced by the silences that occur when the digital-to-analog converters are not fed with data. Therefore, first the entire test signal must be computed, and a single playback of that signal should be used during the test.

The way the recogniser gives output depends very much on the system. Nowadays, word recognition systems are most often shipped as a piece of hardware for a Personal Computer. Proprietary software is often included that allows the user to train the recogniser and to set it up for an application. Unless the controlling computer is the same as the computer that has the recognition system installed, the easiest way to do the assessment automatically is to send the output over a communication line (e.g. an RS-232 port). Some ways for the recognition software to respond are:

- 1. Put recognised words in the keyboard buffer.
- 2. Directly insert the characters into the application (e.g. an editor).
- 3. Return a string on a library call.
- 4. Send the recognised word to a serial communication line.

For the first two possibilities, a clever way has to be devised to be able to send the recogniser's responses to the controlling computer. A simple approach for item 2 is to run a simple terminal emulator program which will automatically send all input to the communication port. The third possibility allows for integration of the recognition system with the controlling computer. The fourth possibility is easiest for a standard set-up with a separate controlling computer.

A stand-alone recogniser is often equipped with a serial communication line to receive commands and to output recognised words. Such systems will fit easily into the general set-up.

The basic assessment procedure in such a set-up is simple, after initialisation and training (see Section 10.5.2):

- 1. Choose a test word, according to the allowed syntax and other defined conditions.
- 2. Instruct the recogniser to "listen".
- 3. Instruct the playback device to play the test word.
- 4. Record the recognition result from the recogniser.

The ESPRIT project SAM has very carefully defined what is called the "Sesam Workstation", as the controlling and speech generating computer system. For this PC-based PC platform many tools have been written, including a recogniser assessment tool "SAMPAC" (see Appendix E. It was developed at TNO-TM (The Netherlands), LIMSI (France), the current developing lab is CSELT, Italy.

10.5.2 Training

It is very important to define the training material for a recogniser. This means specifying

- the database (name, date, supplier),
- the number of speakers,
- the length of the training material (number of training sessions),
- the training conditions (recogniser parameters).

For many purposes, the training and testing sessions are integrated in an overall procedure. For example, the purpose may be tuning the specific recogniser parameters for maximum performance under certain conditions. In this case, it is recommended to split off a certain fraction (typically $\frac{1}{4}$) of the training material and use that for the development test.

In such case, one can actually alter the composition of the development test part of the training material from time to time, in order to prevent "training for the development test material". However, for an actual benchmark of the performance of the system with the final settings of the parameters, unseen test material should be used.

Recommendation 6

For the purpose of tuning parameters of the recogniser, select a part of the training material as test material. Change this selection from time to time.

10.5.3 Test procedure

Once the recognition system has been trained, it can be tested under a set of specified conditions. These conditions may involve adding noise to speech from a database or applying other manipulations to the speech. It is often required that the training has to be adapted to these conditions, e.g. training with noise, etc.

10.5.3.1 Addition of noise

If test conditions involve noise addition, some special precautions have to be made. First it is good to realise that, for a test with added noise, it is possible to have the system trained with or without noise. Because retraining for several noise levels takes additional time, a system is generally trained under one condition of noise, and tested under various conditions. Secondly, it is very important to add the noise *continuously* and *indepen*dently of the speech. This means that the moment at which the noise starts must be independent of the beginning and ending of the words. The reason is that if there is a connection between the start of the noise and the start of a word, this would give a clue to the recognition system where the beginning of the word is, and especially for the assessment of connected word recognition systems this is a major design consideration. In practice, this means that the noise should start a few seconds before the test utterance starts and end only after the recognised words are given. The safest method for stationary types of noise, however, is to have the noise continuously available at the recogniser input.

Recommendation 7

If you test a recognition system under noise conditions, make sure that the noise signal is continuously available, or at least during the period extending from some time before the recognition starts until some time after the recognition has ended.

One way to achieve this is to add analogue noise "outside" the playback device. For simplicity, it would not be a bad idea at all to have a playback device produce the noise during the test, although a disadvantage can be the need for careful adjustment of the SNR. This means that there is less control over the noise level (and thus the signal-to-noise ratio), and one must assure that the levels are correct by measuring the electrical signals (see Chapter 8 for instructions on how to do this). For this purpose the software tool Speech Level Meter (SLM), developed under SAM, can be used (see Appendix E).

The more modern approach to addition of noise is to make a digital addition of noise and test the speech signal. This allows variation of noise within a series of tests quite easily, and signal-to-noise measurements can be performed on a digital level before the mixed signal is fed to the recogniser.

If two signals of comparable level are to be added with limited dynamic range (i.e. the number of bits to represent a sample, often 16 bits), the sample values of both signals have to be divided by 2 first. This is a level

reduction of $6 \,\mathrm{dB}$, while the addition of two uncorrelated signals leads to an increase of the signal level by only $3 \,\mathrm{dB}$. Thus, the total reduction of the level after the signals have been added is $3 \,\mathrm{dB}$. This effect of level reduction due to addition of signals is typically connected to digital signals.

In the case of analogue mixture of equal-level signals, the individual signals have to be attenuated by only 3 dB before they are added, leading to a net null level change. In case the speech level and the noise level are very different, one could in principle use a smaller reduction of the signals, but the reduction of 6 dB is very convenient in practice, because it can be implemented as a simple bit-shift in digital addition.

Recommendation 8

In case you add two signals of comparable level digitally, reduce both signals $6 \, dB$ (a factor 2 in amplitude) before the addition. Be aware of the fact that this will lead to a reduction of level.

If a digital addition of noise is chosen, another point is important: the sample rates of the signals must be the same. This may seem a trivial remark, but in practice this occurs more often than you would expect. Most signal processing software can re-sample a signal digitally, but this may demand resources such as special digital filters. In the SAM project, the program "RESAM" was developed for solving the problem of the dual sample rate standards of 16 kHz and 20 kHz (see Appendix E). This software tool comes with a utility to add digital signals.

In the SAM project and within the NATO research study group RSG 10 several efforts have been undertaken to produce standardised digital noise files. The first product is a CD-ROM entitled "Noise-ROM-0", and contains 24 different noise-like signals, each 4 minutes in length, at 20 kHz sampling rate. The noises vary from wideband reference noises (5 types) to noise from shotguns, cars, aircraft, armed vehicles, etc. The CD-ROM is produced by TNO Human Factors Research Institute in the Netherlands and RSRE Speech Research Unit in England.

The second noise database is called "Noisex", and is distributed on two CD-ROMs. It contains English digits under various calibrated noise conditions. The database has been produced by DRA Speech Research Unit in England.

10.5.3.2 Automatic gain control

Some recognition systems are known to be sensitive to slowly varying level fluctuations. This could be the result of a varying mouth-microphone distance, differences in the speakers emotional state, etc. A solution to this problem is the insertion of an automatic gain control (AGC) in the signal path. However, AGCs tend to have the annoying habit of increasing the gain continuously in periods when there is no speech, until the background noise has reached the required level. At the time the speaker starts talking again, an overload occurs and the gain has to be brought back to a lower level immediately. The solution to this problem is to have the AGC detect silences and not increase the gain during these silences. Although this is in principle possible to do with an analog circuit, nowadays a digital AGC is more convenient. An AGC can be implemented with a digital signal processor board in a Personal Computer.

10.5.4 Scoring the results

Within the SAM project, special software has been developed for scoring the results of a recognition test. The program is called "SAM_SCORE", and is designed for small vocabulary systems (see Appendix E). It is able to determine confidence limits for the performance measures and do a McNemar test on the difference in performance of two recognisers. The program can both be used in batch mode and in interactive mode with a graphical user interface. It can produce output in a DBMS format, which can be used for subsequent analysis software (see below and Appendix H on DBMSs).

10.5.5 Analysis of results

10.5.5.1 Basic statistics, ANOVA

Within SAM, a program called SAMITPRO was developed that can do an Analysis of Variance (ANOVA), cf. Chapter 9, on the test results scored by SAM_SCORE. The technique used is called "iterative proportional fitting". The goodness of the fit is expressed by Pearson's χ^2 . The input of the analysis program is DBMS format. For availability, see Appendix E.

10.5.5.2 Confusion matrices

Based on the confusion matrix, a multidimensional scaling analysis can be performed. For this purpose, the confusion matrix has to be transformed to a symmetric matrix (i.e. $C_{ij} = C_{ji}$), and analysed by multidimensional scaling analysis (HDSCAL) such as developed by Carroll and Chang (1970). Examples of this procedure can be found in Steeneken (1989).

10.6 Experimental design of large vocabulary continuous speech recognition

We will now concentrate on the assessment procedure for speaker independent, large vocabulary, continuous speech recognition systems. At the time of writing, these systems still are laboratory systems, but this may change in the near future.

Unlike the simple isolated word recogniser, these systems generally work off-line, i.e. a whole utterance is fed to the recogniser, and the result may

take a while before it is produced. Often, these systems are assessed in a completely asynchronous way: first all utterances supplied to the recognition system, and later all results are submitted to the scoring program. However, research is also taking place on incremental on-line systems.

The institute with most experience in the assessment of large vocabulary continuous speech recognition systems is the (Defense) Advanced Research Projects Agency (ARPA/DARPA) in the USA. In 1987 this organisation started to organise benchmarking evaluation tests for continuous speech recognition laboratories, which were coordinated and evaluated by NIST (National Institute of Standards). This yearly test has been a great stimulus for the competing laboratories and has proved to be a positive impulse for developing better recognition systems. In the meantime, better training databases have become available, which has also had a positive influence on the results.

This section is heavily based on the ARPA benchmark paradigm, although some experience of the ongoing project SQALE has also been used. Therefore the underlying purpose of assessment is benchmarking. In the ARPA paradigm, systems of various laboratories are evaluated competitively. Before the actual assessment test, training material is defined and distributed (including development test material), and a *dry run* test is performed. After the assessment test, a closed workshop is organised in which the results and benefits of the techniques of the various systems are discussed.

10.6.1 Training material

Contrary to the small vocabulary word recogniser, a large vocabulary recognition system generally uses subword units such as *phones* as acoustic units for recognition. This keeps the number of models to be trained limited (to the number of phone(eme)s in a language, typically 35–50), although training of context-dependent phones (so-called triphones) again increases the number of models (to typically 2000).

This means that the training vocabulary is not necessarily complete with respect to the recogniser's vocabulary. Instead, the recognition system uses a *dictionary* to find the possible ways to pronounce each word in terms of the limited set of phones. The words in the dictionary define the recogniser's vocabulary.

The fact that these systems are designed to recognise *continuous speech* means that they are equipped with algorithms that can segment the input utterance into distinct words. This process of segmentation often can also be used during the training process: this relieves the training databases from giving labelling information on the word boundaries in the speech files.

When organising competitive assessment of various systems, it is important to have carefully defined the training that is allowed. This includes acoustical and language modelling training. In the ARPA paradigm, part of the evaluation test puts fewer restrictions on training material, but demands that this material is available to other participants.

10.6.1.1 Acoustic training

The acoustic training material consists of large databases, with many hours of speech recorded from many people. The most famous training database for American English is the "Wall Street Journal" database (WSJ), with two releases, WSJ0 which contains 84 speakers, and WSJ1, which contains an additional 200 speakers. The total training time is approximately 60 hours. The training sentences come from the Wall Street Journal newspaper. All training sentences have been orthographically transcribed.

It is important that the acoustic training material comes with orthographic transcriptions; without these the material is virtually worthless for training. The size of the material is also relevant; often large vocabulary systems work with models for phone sequences, with up to 3 phones (triphone). This means that the number of models to be trained is quite large, typically 2000. All the models must be trained many times with many different speakers in order to be robust. Up to now, there is no indication that the recognition result as a function of the amount of training material saturates at the available maximum of 60 hours.

10.6.1.2 Lexicon

For training the phone models, some automatic conversion from the orthographic texts to the phones is necessary. This is generally performed by dictionary lookup. Compiling a dictionary is very laborious, and more often than not these dictionaries are considered proprietary information. It is often viewed as part of the recognition system. Recognisers may fall back on a text-to-speech (TTS) system if words in the transcription do not occur in the dictionary (see Chapter 6).

10.6.1.3 Language model

An essential part of the large vocabulary continuous speech recognition system is the language model. It represents the machine's knowledge of the language it is supposed to recognise. Because the recognisers often have a probabilistic approach to the acoustic modelling, a probabilistic language model fits in perfectly. There are many ways to implement a probabilistic grammar (see Chapter 7, but the most widely used is the *n*-gram grammar. In an *n*-gram grammar, the probability that word w_n follows a sequence of words $w_1, w_2, \ldots, w_{n-1}$ is defined. The number of possible combinations of *n* consecutive words is V^n , where *V* is the vocabulary size. For n = 3and $V = 20\,000$, the number of trigrams that must be known is astronomic, 8×10^{12} . Apart from storage problems, it would require enormous amounts
of text just to see all combinations at least once. Therefore, techniques have been developed to deal with this problem. One of these is the backoff principle. In this technique, an untrained *n*-gram is expressed as the product of a *back-off probability* and the (n-1)-gram of the final n-1words. The back-off probability depends on the *first* n-1 words. This process can be continued recursively, up to the unigram probabilities.

The *n*-gram probabilities and back-off probabilities have to be trained with large amounts of text. A common source for benchmark evaluation is newspaper texts, but in principle the domain should match that of the application. If the application is dictation of law texts, a good choice for training texts are (electronic versions) of law books. Getting these texts electronically might be difficult, and in all cases copyrights have to be respected¹ In order to give the reader an idea of the text sizes: in the November 1993 ARPA benchmark evaluation the standard language model was trained with 37 million words of WSJ text, in the December 1994 evaluation the language model training material increased to 237 million words from 5 sources. Generating *n*-grams from texts has been standardised by Carnegie Mellon University. They have made a toolkit which is freely available for research purposes.

The language model can be precompiled. In fact, in the ARPA benchmark evaluations a trigram language model is shipped with the training material that has a precompiled form. This model was built by Doug Paul, then at MIT (Cambridge, MA), and the format of this language model has become the *de-facto* standard. It is very simple in structure. The file is a simple text file, starting with some comments. The header ends with the keyword \data, after which the keyword \ngrams: starts an n-gram block. In the following lines, each line specifies the n-gram probability, the n words, and a back-off probability. The probabilities p are given as ¹⁰ log p. For instance, in the block containing 2-grams, one may find a line like

which should be interpreted as follows: the probability that the word *dollars* occurs, given the fact that the previous word is *million*, is $10^{-0.1924} = 0.6421$. If a trigram "million dollars w_3 " is not specified in the file, try to use the bigram probability "million w_3 " and correct for backing-off with an extra factor $10^{-0.0922} = 0.8088$. (For instance, the word w_3 might be "left", and the training combination "million dollars left" might not have occurred in the training texts. If in the recognition process the probability for this combination must be estimated, this back-off procedure is used.)

 $^{^{1}}$ For problems pertaining to written language (text) corpora, the results of the EA-GLES Working Group on Corpora should be consulted.

10.6.2 Development test

It is very likely that a large vocabulary recognition system has not yet been fully tuned for the application. In order to make it possible for the system to be tuned, some *development test material* is required. In principle this material should be very similar to the assessment test material, but of course essential details should differ. The utterances, for instance, should not be the same, and if the system is speaker independent, there should be no overlap in speakers. If no special recordings have been made for development testing, one could split off a fraction of the training material and use that for development testing.

10.6.3 Dry run

In the ARPA style of competitive evaluation, a special test before the actual evaluation is defined, which is identical to the actual evaluation test, but results are not made public. This test is in fact a rehearsal for the evaluation test and includes all actions to be performed by the organiser of the test and the participating laboratories. Again, the test material should be different from the assessment test (different speakers, utterances, etc.) but all recording conditions, file formats, file naming, media types, submission protocols, etc. should be as much the same as possible.

In the EC project SQALE, where assessment of 4 languages with 4 different recognition systems is performed, the functionalities of the dry run and that of the development test are combined into one single test. This implies that the period between the development test or dry run and the evaluation test must be long enough to allow for system development.

10.6.4 Test material selection

The selection of the test sentences requires some attention. If the assessment purpose is purely benchmarking for a very specific application, one could use the "representative database" approach and select utterances randomly from a source of possible utterances. For large vocabulary recognition systems, the set of possible utterances is virtually unlimited, however, and this means that a random selection will actually be from a pre-selected set of utterances. In the ARPA paradigm, this pre-selection is formed by selecting paragraphs randomly from a specified period of time from specified newspapers. For the "20 k open vocabulary test" of the 1993 evaluation, the paragraphs were pre-filtered to contain only words from the most frequent 64 k words in the WSJ. For the "5 k closed test" the words were restricted to be in the top most 5000 words of the frequency-sorted list.

If the purpose of the assessment is a little more diagnostic, the following considerations might be made. From previous experiments with continuous speech recognition assessment it is known that very basic parameters, such

as gender and sentence perplexity, influence the recognition result. Because the diversity in speakers is responsible for another important part of the variability in recognition score, it is wise to balance the earlier mentioned parameters over the speaker. For the parameter gender, this has the logical consequence of having as many male speakers as female speakers, which is contrary to the representative purpose of assessment, where the gender ratio of the users of the foreseen application should be reflected; for fighter pilots, the ratio could lean towards male speakers. The number of out-ofvocabulary words in the set of sentences should be kept constant across the speaker. The distribution of sentence perplexity should be more or less the same across speakers. Thus one can obtain diagnostic information on what part of the variability is due to speaker variation and what part is due to other factors, such as gender, perplexity and fraction of out-of-vocabulary words.

With large vocabulary recognition systems one is generally not able to test all words in the vocabulary, as in assessment of small size vocabulary words recognisers. This is not really necessary, because generally the systems are phone-based and it is more important to cover all phones in representative amounts in the assessment. This allows the size of the test to be restricted to typically 20 speakers uttering 15 sentences of approximately 20 words on the average. This corresponds to typically 6000 words in the test and a multiple of this number of (context dependent) phones. A large vocabulary system typically uses 500–3000 models, and this size of test more or less covers the phones which are modelled.

In order to cover as many words (or context dependent phone models) as possible, in the ARPA style of benchmarking assessment all speakers utter distinct sentences. From a more diagnostic point of view, it would be ideal to have all speakers utter the same sentence, in order to be able to distinguish variability caused by the speakers and variability caused by sentences. This is practically impossible, however, because this would both need many recordings and a long recognition time. One way to overcome this problem, as is done within SQALE, is to divide the evaluation test into two parts. One part consists of the "classic" 20 speakers, 10 sentences per speaker, all sentences unique. The other part consists of extra sentences for variance estimation. This part has few sentences (typically 3) uttered by 10 different speakers, where the sentences are the same across the speakers. Additionally, typical 6 speakers utter the same sentence 5 times (each speaker a different sentence). These replicas allow the estimation of the variance within one speaker for the same sentence. Although these utterances are different from the ones used in the first part of the assessment test, they may shed light on the different contributions to the variance.

10.6.5 Evaluation protocol

If one organises a competitive benchmarking evaluation among several speech recognition laboratories, an evaluation test protocol should be kept. In this protocol, the coordinator should specify all details concerning the test. This is necessary, because for such an evaluation the recognition is not performed at the location of the organising institution, but at the sites of the laboratories themselves. Items that should certainly be defined in such a protocol are:

- Dates and times when the evaluation test begins, when the results are due, when the adjudication period ends, when the official results will be published.
- A description of how the recognition results are to be submitted to the organising institute, file formats, file structure, etc.

10.6.6 Scoring method

For continuous speech it is not so trivial how the assignment of "deletions", "substitutions" and "insertions" should be made. The process by which this is carried out is called *alignment*. If the recogniser's segmentation is available, i.e. if the times of the starting and ending of each recogniser word are available, this alignment can be done in a way comparable to the isolated word recogniser assessment.

Generally, such labelling information is not available in the recognition output. In this case, the alignment process uses a dynamic programming algorithm to minimise the misalignment of two strings of words (symbols), the reference sentence and the recognised sentence. The alignment depends on the relative weights of the contributions of substitutions, insertions and deletions. Hunt (1990) discusses the theory of word-symbol alignment and analyses some experiments on alignment.

NIST has developed freely available software for analysis of continuous speech recognition systems. It basically consists of two parts: an alignment program and a statistics package.

The alignment can be performed both on word level and on the phone level (so-called phonetic alignment) if the dictionary is available. It is a standard alignment procedure and is therefore recommended for competitive assessment.

The software was developed for the ARPA evaluations, but it has been designed to make the programs generally applicable. The alignment program generates a (binary) file with all alignment information, which can be printed by another utility in various levels of detail. Overall results can be compiled, as well as results on a per-speaker level. The statistics program can pairwise compare the results of different recognition systems and decide whether or not the difference in performance is significant. This is done using four different statistical tests.

11 Assessment of speaker verification systems

11.1 Presentation

Until about a century ago, under most circumstances to be within range of the voice was to be within range of the eyes, and the face or silhouette were no doubt much more important to humans than the voice as a means of identifing each other, with the exception of special cases such as the blind. With the development of telecommunications and acoustic recordings, the need for speaker recognition has become more important.

The basis for automatic speaker classification and recognition is that in addition to the linguistic message, the human voice conveys a lot of paralinguistic information about the speaker, i.e. the "encoder". These factors of variability are well-known obstacles to speech recognition, as they increase the variability of the speech signal.

The main sources of a speaker's specificity are the physiological configuration of his speech production organs, his neuro-motor control of these organs, and his internal speech pattern prototypes. In practice, there may exist more or less systematic correlations between these factors and some of the speaker's characteristics, such as his sex, age, health conditions, mood, regional, cultural, educational background, possible foreign accent, and the language he is speaking.

In this chapter, we address a class of pattern recognition problems where the goal is to classify a speech pattern according to some characteristics of the speaker who uttered it. We recommend the general term *speaker classification* to denote such problems.

11.1.1 Speaker classification tasks

- If the goal is to decide whether a given speech utterance was uttered by a male speaker or a female speaker, this particular problem of speaker classification can be referred to as *sex identification*.
- When the goal is to classify a speaker within an age group, from a spoken utterance, the problem can be called *age identification*.
- Some health professionals are interested in detecting pathologies using voice samples (for instance, vocal cord disfunctionings). This problem of pathology identification is a particular case of *health state identification*.¹
- Any task that would consist in determining whether a speaker is angry, sad, stressed, calm, happy, relaxed, etc. would resort to *mood identification*.²

¹This concept could be extended to the characterisation of voices modified by external temporary factors that affect speech production, such as alcohol for instance.

 $^{^{2}}$ It is essential to underline here that this sort of problem is not solved yet, and will

- We will understand by the term *accent identification* any process consisting of determining some aspects of the sociological background of the speaker. The most realistic is certainly to try to identify regional accent for a native speaker, or linguistic origin for a non-native speaker.
- For some applications, it is necessary to classify a speaker with respect to one of several categories, the characteristics of which cannot be expressed in objective terms³ This task can be covered under the general term of *speaker cluster selection*.

In the special case where the goal is to identify in which language a given speech utterance has been produced, we recommend using the term *spoken language identification* instead of the usual expression of *language identification*, as the latter can be confused with *written language identification*.

Finally, if the task consists in finding information about the identity of the speaker from a speech signal, it is classically designated as *speaker recogni*tion.

For speaker classification and recognition tasks, a general distinction must be made between *identification* and *verification*. While *identification* consists in finding to which class or speaker a speech utterance is most likely to belong, *verification* aims at validating or dismissing the hypothesis that the utterance pertains to a given class or speaker.

Examples of *speaker class identification* are given above. For *speaker class verification*, a typical problem of *age verification* would consist in checking whether a speaker is an adult or not, and *spoken language verification* would aim at checking whether an utterance was pronounced in a given language (the expected language of an application, for instance).

In the rest of this chapter, we will mainly focus on speaker identification and verification. However, most concepts are easy to generalise to other speaker classification problems.

11.1.2 General definitions

The following general definitions are directly based on Atal (1976):

Speaker classification: any decision-making process that uses features of the speech signal to determine characteristics of the speaker of a given utterance. Speaker recognition: any decision-making process that uses features of the speech signal to determine information on the identity of the speaker of a given utterance.

Speaker class identification: any decision-making process that uses some features of the speaker signal to determine the class the speaker of a given

probably never be. In particular, lie detection from the speech signal is not considered as a realistic research area.

 $^{^3}$ For instance, some speech recognition systems use models of speech units that have variants across several speaker clusters. These clusters may be obtained in an unsupervised manner, and it is usually impossible to find *a posteriori* an objective attribute that would qualify each cluster.

utterance belongs to.

Speaker class verification: any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance belongs to a given class.

Speaker identification: any decision-making process that uses some features of the speech signal to determine who the speaker of a given utterance is.

Speaker verification: any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance is a particular person, whose identity is specified.

Spoken language identification: any decision-making process that uses some features of the speech signal to determine what language is spoken in a given utterance.

Spoken language verification: any decision-making process that uses some features of the speech signal to determine whether the language spoken in a given utterance is a particular language.

11.2 A taxonomy of speaker recognition systems

Any speaker classification or recognition system functions in 3 modes:

- A *training mode*, during which speaker or speaker class models are built, estimated and stored in the system, as reference patterns. Alternative terms: learning mode, registration, enrolment, subscription, etc.
- A *test mode*, during which the system performs the recognition of an utterance to be identified (or verified). Alternative terms: recognition mode, trial mode, operating mode, etc.
- An *untraining mode*, during which a speaker or speaker class model is removed from the list of reference patterns. Alternative terms: unlearning mode, unregistration, subscription cancellation, etc.

Naturally, a given speaker cannot be identified or verified in the test mode before he has been registered during a training phase. However, training is usually incremental in several ways. Firstly, a speaker recognition system is usually operational before all possible speakers are referenced, and new users can be added while the system is already functioning in test mode for those who are registered. Secondly, some system implementations use speech material obtained during a test phase to update training references (when the recognition decision is judged reliable enough).

Unlike training and testing, untraining does usually not require active participation of the speaker being unregistered.

11.2.1 Task typology

11.2.1.1 Speaker identification versus speaker verification

As we briefly mentioned in the introduction, speaker recognition covers two different areas: on the one hand, speaker identification, on the other hand,

speaker verification. As Doddington (1985) describes it, the goal of a *speaker identification* task is to classify an unlabelled voice token as belonging to one of a set of n reference speakers, whereas the *speaker verification* task is to decide whether or not the unlabelled voice belongs to a specific reference speaker.

Speaker identification is therefore a 1 out of *n* decision, in the case of *closed*set speaker identification, the result of which is an *identity assignment* to an *applicant speaker*. However, in practical applications, *open-set* speaker identification requires an additional outcome of *rejection*, corresponding to the possibility that the unlabelled speech token does not belong to any of the *registered speakers*. In such circumstances, the applicant speaker is called an *impostor*.

Speaker verification can be viewed as a particular case of open-set speaker identification, corresponding to n = 1. The speaker verification system takes a test voice sample and a *claimed identity* as input, and returns a binary decision: *acceptance* if the applicant speaker is considered to be the *genuine speaker* or *rejection* if he is considered to be an *impostor* (as regards the claimed identity).⁴

Conversely, open-set speaker identification can be understood as a step of closed-set speaker identification, followed by a step of speaker verification, the latter using the identity assigned by the former, as the claimed identity.

11.2.1.2 Related tasks

Beyond this major distinction between identification and verification, other related tasks can be mentioned.

- *Speaker matching*, that is, to choose a speaker in a closed-set of references which is most similar to a current speaker, even though it is known in advance that the applicant speaker is not a registered speaker. This appears to be a particular case of speaker cluster selection, where each cluster consists of one speaker only.
- Speaker labelling, when the identity of speakers taking part in a conversation are registered, and the goal is to localise when their successive interventions begin and end.⁵
- *Speaker alignment*, when the identity and order of speakers taking part in a conversation are known and the goal is to localise when each of their interventions begins and ends.
- Speaker change detection, when the goal is to detect a change of speaker along a speech stream.

⁴Note here that the term *impostor* covers two slightly different concepts: a non-registered speaker in identification, and a speaker claiming a false identity in verification. ⁵Including a possible outcome of *none of the registered speakers*, in case of open-set labelling.

11.2.1.3 Types of error

For closed-set speaker identification, a *misclassification* error occurs when a registered speaker is mistaken for another registered speaker (the *mistaken speaker*).

For speaker verification, two types of error must be distinguished: *false* rejection when a genuine speaker is rejected and *false acceptance* when an impostor is accepted as the speaker he claimed he was (the *violated speaker*). For open-set identification, the three types of error can occur. Usually, misclassifications and false acceptances are considered as equally harmful, and therefore merged together. However, these two types of error may not have the same consequences in some practical applications.

11.2.2 Levels of text dependence

Another feature which is classically used to specify a speaker recognition system is its level of *text dependence*, i.e. the constraints on the linguistic material imposed on a test utterance. A main distinction is conventionally set between *text-dependent* systems and *text-independent* systems. Though this basic distinction is not accurate enough to cover the range of practical possibilities, below we give a definition of these two terms according to the usage found in the literature. To simplify, in text-dependent systems, the linguistic content of the training and test material are totally identical, while in text-independent systems test utterances vary across trials (at least in terms of word order).

However, a deeper study of the various strategies used in practice shows that at least five levels of text dependence should be distinguished. Two of them resort to text-dependent approaches, but can be opposed to the use of either a *personal password*⁶ or a *common password*.⁷ The other three can be viewed as several variants of text-independent approaches, using either fixed words in a random order (*fixed-vocabulary* systems⁸), a specific linguistic event, wherever it occurs (*event-dependent* systems⁹), or a completely unrestricted text (*unrestricted text-independent* systems¹⁰).

11.2.3 Interaction mode with the user

For text-dependent systems, the speech material that the user must pronounce in front of the system is *a priori* determined during the registration. While common-password systems have absolutely no flexibility in the choice of the linguistic material, personal-password systems can enable some text

 $^{^{6}\,\}mathrm{For}$ instance, its own name, or a personal identification number.

⁷ For instance, the voice request for a given protected service.

 $^{^{8}\,\}mathrm{For}$ instance, having the applicant speaker pronounce a new sequence of digits for each trial session.

 $^{^9}$ For instance the vowel [i], a nasal sound, the word /dog/, ...

 $^{^{10}\,\}mathrm{However},$ such systems may be language dependent.

customisation, in particular the possibility for the registered user to change his voice password.

For text-independent systems, at least three modes of interaction with the user can be distinguished: *text-* and *voice-prompted* systems impose the (unpredictable) linguistic material to the user, whereas *unprompted systems* use totally spontaneous speech. In parallel with speaker recognition, prompted systems explicitly or implicitly perform some kind of speech recognition in order to check that the applicant speaker has really uttered what he was asked to say.

11.2.4 Definitions

Applicant speaker: The speaker using a speaker recognition system at a given instant. Alternative terms: current speaker, test speaker, unknown speaker,¹¹ customer, user, etc.

Registered speaker: A speaker who belongs to the list of registered users for a given speaker recognition system.¹² For speaker classification systems, we propose the term *conform speaker* (conformant speaker) to qualify a speaker who belongs to one of the classes of speakers for a given speaker classification system.¹³ Alternative terms: reference speaker, valid speaker, authorised speaker, subscriber, client, etc.

Genuine speaker: A speaker whose real identity is in accordance with the claimed identity. By extension: a speaker whose actual character and claimed class are in accordance.¹⁴ Alternative terms: authentic speaker, true speaker, correct speaker, etc.

Impostor (speaker): In the context of speaker identification, an impostor is an applicant speaker who does not belong to the set of registered speakers. In the context of speaker verification, an impostor is a speaker whose real identity is different from his claimed identity. Alternative terms: impersonator, usurpat,¹⁵ etc. For speaker classification tasks, this concept is better rendered by the term: discordant speaker.¹⁶

Identity assignment: Decision outcome which consists in attributing an identity to an applicant speaker, in the context of speaker identification.

¹¹This term can be ambiguous in certain contexts, as it may also be understood as a speaker who is unknown to the system. Though it is frequently found in the literature, we do not recommend using it.

 $^{^{12}}$ Usually, a speaker who is entitled to use the facilities, the access of which is restricted by the system. 13 For instance, for a spoken language identification system that discriminates between

¹³For instance, for a spoken language identification system that discriminates between languages spoken in Switzerland, a *conform speaker* (conformant speaker) is a speaker who speaks either German, French, Italian or Romansch, but not some other language the system does not expect.

 $^{^{14}\,\}mathrm{For}$ instance, a female speaker claiming that she is a female speaker, in sex verification.

¹⁵Both terms are very rarely used.

¹⁶For instance, a child claiming that he is an adult, in age verification.

For speaker classification, the term *class assignment* should be used instead. *Acceptance*: Decision outcome which consists in responding positively to a speaker (or speaker class) verification task.

Rejection: Decision outcome which consists in refusing to assign a registered identity (or class) in the context of open-set speaker identification (or classification), or which consists in responding negatively to a speaker (class) verification trial.

(Speaker) misclassification: Erroneous identity assignment to a registered speaker in speaker identification.

False (speaker) rejection:¹⁷ Erroneous rejection of a registered speaker or of a genuine speaker in open-set speaker identification or speaker verification respectively.

 $False (speaker) \ acceptance:^{18} \ Erroneous \ acceptance \ of \ an \ impostor \ in \ openset \ speaker \ identification \ or \ in \ speaker \ verification.$

Mistaken speaker: The registered speaker owning the identity assigned erroneously to another registered speaker by a speaker identification system. *Violated speaker*: The registered speaker owning the identity assigned er-

roneously to an impostor in open-set speaker identification system. The registered speaker owning the identity claimed by a successful impostor, in speaker verification.

Text-dependent speaker recognition system: A speaker recognition system for which the training and test speech utterances are composed of exactly the same linguistic material, in the same order (typically, a password).

Text-independent speaker recognition system: A speaker recognition system for which the linguistic content of test speech utterances varies across trials. *Personal-password* speaker recognition system: A text-dependent speaker recognition system for which each registered speaker has his own voice password.

Common-password speaker recognition system: A text-dependent speaker recognition system for which all registered speakers have the same voice password.

Fixed-vocabulary speaker recognition system: A text-independent speaker recognition system for which test utterances are composed of words, the order of which varies across speakers and sessions, but for which all the words were pronounced at least once by the speaker when he registered to the system.

Event-dependent speaker recognition system: A text-independent speaker recognition system for which test utterances must contain a certain linguistic event (or class of events) while the rest of the acoustic material is discarded. This approach requires a preliminary step for spotting and localising the relevant events.

¹⁷Sometimes called type-I error.

¹⁸Sometimes called type-II error.

Unrestricted text-independent speaker recognition system: A textindependent speaker recognition system for which no constraints apply to the linguistic content of the test speech material.

Text-prompted speaker recognition system: A speaker recognition system for which, during the test phase, a written text is prompted (through an appropriate device) to the user, who has to read it aloud.

Voice-prompted speaker recognition system: a speaker recognition system for which, during the test phase, the user has to repeat a speech utterance, which he listens to through an audio device.

Unprompted speaker recognition system: A speaker recognition system using totally spontaneous speech, i.e. for which the user is totally free to utter what he wants,¹⁹ or for which the system has no control over the speaker.²⁰

11.2.5 Examples

In this section, we give examples of well-known speaker recognition systems which can be found in the literature, in order to illustrate the taxonomy described above.

11.2.5.1 Text-dependent systems

Among examples of text-dependent systems, the *Bell Labs system*, reported by Rosenberg (1976) and improved by Furui (1981) is tested by the latter under the following protocol: "Several [six] kinds of utterance sets were used to evaluate [the] system ... Two all-voiced sentences were used in the recordings. The males used the sentence, 'We were away a year ago' and the females used the sentence, 'I know when my lawyer is due.'" (Furui 1981, p. 258)

The first five utterance sets are composed of male speakers, while the last one is composed of female speakers. Performances are reported for speaker verification experiments on each set. Following our terminology, these experiments simulate a *common-password text-dependent speaker verification system*. Here, the password is an entire sentence. As Rosenberg notes to justify the use of a text-dependent system in practical applications: "For many applications, the speakers are expected to be cooperative so that a prescribed text is perfectly feasible." (Furui 1981, p. 259)

The use of a prescribed text has also the advantage that it does not need any prompting, but the drawback that it may be forgotten by the user. As discussed in a next example, a second drawback of text-dependent systems is the possibility for impostors to use pre-recorded speech.

 $^{^{19}\,{\}rm Here,}$ a further distinction could be made between language~dependent and language~independent systems.

 $^{^{20}\,{\}rm For}$ instance, in for ensic applications, the speaker may not be physically present, or may not be willing to cooperate.

As an example of *personal-password text-dependent speaker verification*, one can mention a new service offered by the American telephone operator *SPRINT*. For this service, the user must speak his telephone card number through the phone, in order to have his home bill charged directly for the call he is willing to make. The system identifies the claimed customer by recognising the sequence of digits, and then verifies, on the same sequence of digits, the match between the actual user and the assumed customer. Here, the sequence of digits has a double function: a means of customer identification, and a personal voice password for speaker verification.

11.2.5.2 Fixed-vocabulary systems

Another very popular speaker verification systems was developed by Doddington at Texas Instruments, in the early 70s. Here follow excerpts of the description given by the author (Doddington 1985, p. 1661):

To use the system an entrant first opens the door to the entry booth and walks in, then he identifies himself by entering a user ID into a keypad, and then he repeats the verification phrase(s) that the system prompts him to say. If he is verified, the system $[\ldots]$ unlocks the inside door of the booth so that he may enter into the computer center. If he is not verified, the system notifies him by saying "not verified, call for assistance".

Verification utterances are constructed randomly to avoid the possibility of being able to defeat the system with a tape recording of a valid user. An simple four-word fixed phrase structure is used, with one of sixteen word alternatives filling each of the four word positions (see Table 11.1).

 Table 11.1: Verification Phrase Construction for the TI Operational

 Voice Verification System (after Doddington)

GOOD	BEN	SWAM	NEAR
PROUD	BRUCE	CALLED	HARD
STRONG	JEAN	SERVED	HIGH
YOUNG	JOYCE	CAME	NORTH

An example verification utterance might be "Proud Ben served hard". These utterances are prompted by voice. This is thought to improve verification performance by stabilising the pronunciation of the user's utterance.

Therefore, the TI system turns out to be a voice-prompted fixed-vocabulary speaker verification system, the claimed identity being input as a personal

identification number on a keypad. Doddington's excerpt illustrates well the motivations behind the voice-prompted fixed-vocabulary approach: the relative randomness of the verification utterances protects against impostors using pre-recorded speech, while the use of voice prompts tends to control the reproducibility of the user's pronunciation. However, it must be noted that voice-prompting may also neutralise some of the speaker characteristics (in particular prosodic factors), owing to an unconscious mimicry of the prompt. At the same time, text-prompting has the drawback of requiring a specific device, such as a screen, which is not always possible to implement.²¹ The experiments reported by Soong et al. (1987) where sequences of digits are used for speaker verification is another example of a fixed-vocabulary system.

11.2.5.3 Unrestricted text-independent systems

Unrestricted text-independent speaker recognition is usually considered as desirable for several reasons. Even if the user does not have to take the initiative in producing the text, prompted systems are less likely to be defeated by a recorded voice, as the linguistic material is virtually unpredictable. For unprompted systems, identification or verification can take place unobtrusively, during a telephone transaction, for instance. Moreover, unprompted approaches do not require the speaker to be actively cooperative.

Here is the general structure of a *text-* (or *voice-*) *prompted unrestricted text-independent system*, as described by Furui (1994), p. 7:

The recognition system prompts each user with a new key sentence every time the system is used, and accepts the input utterance only when it decides that the registered speaker has uttered the prompted sentence $[\ldots]$ This method not only can accurately recognise speakers but also can reject utterances whose text differs from the prompted text, even if it is uttered by the registered speaker.

[During registration], since the text of training utterances is known, these utterances can be modelled as the concatenation of [speaker-independent] phoneme models, and these models can be automatically adapted [to the new registered speaker]. In the recognition stage, the system concatenates phoneme models according to the prompted text [i.e. a speaker-specific model and a speaker-independent model]. If the likelihood²² of both speaker and text is high enough, the speaker is accepted as the claimed speaker.

 $^{^{21}}$ On the telephone, for instance.

²²The author refers to a probabilistic model, namely a Hidden Markov Model.

Note here that the fundamental difference between the system described above and a fixed-vocabulary system is the use of subword speech units (here, phonemes) which allow to construct speaker-specific models of test words (or sentences) which were not pronounced during the registration phase. Note also the use of an explicit step of *speech* recognition.

In opposition to prompted systems, here is one example of an experiment in unprompted speaker recognition, as reported by Gish et al. (1986), p. 865, concerning the ISIS system from BBN:

We wish to identify an unknown speaker, from an utterance, $[\ldots]$ knowing that the utterance was made by one of a set of M possible speakers. We have available training data for each of the M speakers that consists of speech from one or more telephone calls, all distinct from the test telephone call. The text of all utterances is assumed to be unknown.

Here, the protocol described is *unprompted unrestricted text-independent* closed-set speaker identification. Note also the *multi-session* character of the experiment, i.e. that the training and test material have been recorded through different channels, probably on different days.

11.3 Influencing factors

Many factors influence the performance of a speaker recognition system. In this section, we list the most obvious ones and stress their role on the general performance of speaker recognition systems. In particular, we discuss which are the most relevant aspects which should be explicitly taken into account and reported on when defining and describing an evaluation protocol, and how they should be handled and expressed.

11.3.1 Speech quality

A major factor influencing the performance in speaker recognition is *speech quality*, in particular the bandwidth of the recording, the location of the speech acquisition system (anechoic chamber, studio, office, street, ...), the type of microphone used, and the transmission channel (direct recording, telephone transmission, radio transmission, type of signal coding, ...).

In practice, it is essential that the description of speech quality reports accurately the general characteristics, but also underlines what the variable factors and those that remain constant are between several trials of the same user.

For instance, speaker recognition through the telephone is generally more difficult than speaker recognition using studio recordings. But two evaluations using telephone speech may give different results if for the first one each speaker always calls from the same place using the same telephone, while for the second one he is asked to call from different locations. Naturally, the latter is likely to give better results than the former.²³

11.3.2 Temporal drift

The characteristics of a voice vary over time, depending on how tired the speaker is, how stressed he is, what mood he is in, whether he has a cold or not, etc. Moreover, it was often noted that the behaviour of users changes while they are getting accustomed to a system. These trends can be gathered under the term *temporal drift*.

Temporal drift usually affects significantly the performance of a speaker recognition system. Intra-speaker variability within a single recording *session* is usually much smaller than *inter-session* variability. In practice, performance levels deteriorate significantly a few days, or even a few hours, after registration, as compared to those obtained with *contemporaneous speech*, i.e. when test utterances are pronounced immediately after the training phase is terminated. A partial solution to temporal drift consists in using training material which is gathered over several sessions: as the collected data are more representative of the intra-speaker variability over time, more robust speaker models can be built. However, this approach makes the registration process heavier.

When the targeted application is intended to operate along time, it is necessary to design an evaluation experiment for which test material was recorded in several sessions, separated from each other by at least one day, and covering a reasonable time-span (at least a month).²⁴ When multi-session recordings are available, the training material should be chosen so that it corresponds to the first recording session (or sessions, for multi-session training). Conversely, the material of a given session should never be split between training and testing,²⁵ as this would lead to an unrealistic protocol. When these constraints are fulfilled, the number of training and test sessions, and the timespan covered by both phases should be explicated. Note that the number and timespan of training sessions has an influence on performance levels and on the user acceptability, whereas the number and timespan of test sessions have only an impact on the statistical validity of the evaluation results.

11.3.3 Speech quantity and variety

Another fundamental factor having an effect on performance level is *speech quantity*, i.e. the amount of speech material used during the registration phase, and during each user trial phase. Usually, the level of performance

 $^{^{23}\,\}mathrm{In}$ fact, the system may be more efficient in recognising the hand set rather than the speaker.

 $^{^{24}}$ One year would be ideal, for a good representativity of weather influence.

 $^{^{25}}$ Except, of course, when the evaluation is carried out on contemporaneous speech.

increases with speech quantity, but stabilises after a certain amount.²⁶ As the user's acceptability generally drops when the training phase (and *a fortiori* the test phase) lasts too long, a compromise has to be found when the performance improvement is judged marginal.

When registration consists of several training sessions, it is the *total training speech quantity*, i.e. the *speech quantity per training session* multiplied by the number of training sessions, which has the clearest impact on performance. When tests are carried out with speech material from several sessions, the relevant factor is the *speech quantity per test session*.

It may also be important to distinguish the quantity of speech *uttered* by the speaker from the quantity of speech *actually used* for training and testing. This distinction is mostly appropriate for event-dependent systems and mainly for diagnostic analysis. In this case, the (average) *proportion* of speech actually used can be given for information, for the training mode and for the test mode. However, the *absolute* quantity of used speech should *not* be reported, in order to avoid ambiguities.

A related factor to speech quantity is *speech variety*: for a given speech quantity, it is usually more efficient to cover a wider range of linguistic phenomena.²⁷ Lacking a universal quantitative measure of linguistic coverage, a qualitative description of the linguistic material is the only way of specifying this aspect.

11.3.4 Speaker population size and typology

In this section, we indicate in what way, and to what extent the population composition, in terms of size and typology, can affect the performance of a speaker recognition system, and how it should be taken into account when designing an evaluation experiment.

When the goal is closed-set speaker identification, it is clear that the complexity of the task increases with *n*, the *registered speaker population size*. However, the *proportion of men and women* in the population also has a direct influence, as same-sex confusions are usually much more likely than cross-sex errors.²⁸ If additional geographical, physiological, and even psychological and sociological information seems particularly relevant or clearly specific of the tested population, the experimenter should be aware of it and explicate it.

With respect to concerns *speaker verification*, the level of performance does *not* depend on the registered speaker population size, as for each trial, the complexity of the task corresponds to an open-set speaker identification

 $^{^{26}\}mathrm{See}$ for instance Soong et al. (1987) for an illustration.

 $^{^{27}\}mathrm{See}$ again Soong et al. (1987) for an illustration.

²⁸ In other words, it is intrinsically easier to identify 1 speaker among 10 than to identify 1 among 1000, and it is intrinsically easier to identify 1 male speaker among 1000 adult speakers rather than among 1000 male speakers.

with n = 1. A large representative population of registered speakers will only guarantee a higher statistical validity of evaluation results, whereas general conclusions will be less reliable with a small specific population.

However, a relevant issue for speaker verification (and open-set identification) is the number and typology of *pseudo-impostors*,²⁹ i.e. speakers used to model impostors during the registration phase. With more pseudoimpostors, the modelling of imposture is usually more accurate. The way pseudo-impostors are selected, and in what way they differ from authorised users is also essential.

In general, each registered speaker has a corresponding *impostor model*, which represents real impostors who could claim his identity. The impostor model can be *common* to all registered speakers, or *specific* to each authorised user, if the pseudo-impostor population varies across subscribers. Pseudo-impostors can be chosen within the population of registered speakers, or originate from an external population. We will use the term *pseudo-impostor bundle*³⁰to refer to the group of speakers who have been used to build the impostor model of a given registered speaker.

From a practical point of view, when impostor models are built from other registered speakers, the recording burden is lighter, but the impostor models may be less representative of imposture in general. If an additional population of external speakers is used, the *number of additional pseudo-impostors*, their population typology, as well as the speech quantity and number of sessions required from each of them should be specified.

Incidentally, for the evaluation of a speaker verification system, a test impostor should not be part of the pseudo-impostor bundle of the speaker he is claiming to be, as the real rejection abilities of the system may be overestimated otherwise. On the other hand, there is no objection to having a registered speaker belong to his own pseudo-impostor bundle, as is the case when the whole registered population is used to build a common impostor model.

11.3.5 Speaker purpose and other human factors

The motivation for which a speaker is using a system also influences considerably its performance profile. We first describe a possible typology of applicant speakers as regards their objectives. Then we mention other relevant human factors.

When the user's goal is conformant with the purpose of the system, a *cooperative (registered) speaker* can be defined as an authorised applicant who is willing to be identified or as a genuine speaker who intends to be verified positively. Their counterpart in the impostor population would be a *well*-

 $^{^{29}{\}rm Sometimes}$ called background speakers.

³⁰New term that generalises the concept of *cohort*.

intentioned impostor, i.e. an impostor having the goal of being rejected.³¹ When the user's goal and the system's purpose are inverse, an uncooperative (registered) speaker knows that he is being verified but wants the system to reject him.³² For instance, an uncooperative speaker is likely to use natural or artificial voice masking in order to remain anonymous. In contrast, an intentional impostor has the clear goal of being identified or verified though he is not registered (violation), or to be identified as somebody else (usurpation).

Here, a distinction must be made among intentional impostors depending on whether they previously have or have not been in contact with the voice of the authentic user whose identity they are claiming. We propose the term *acquainted impostor* to qualify an intentional impostor who has some knowledge of the voice of the authorised speaker, as opposed to *unacquainted impostors*, when the impostor has never been in contact with the genuine or authentic user. The degree of success of an acquainted intentional impostor will ultimately depend on his *imitation* skills.

The term *casual impostor* is often used to qualify speakers who are used as impostors in an evaluation, but who were not recorded with the explicit instruction to try to defeat the system. In the same way, the term *casual registered speakers* can be used to refer to a population of registered speakers who have not received an explicit instruction to succeed in being identified or verified positively.³³

Here again, variants appear, depending on the way the experimenter chooses the claimed identity of a casual impostor in a verification experiment. A casual impostor can be tested against all registered users systematically,³⁴ against all other registered speakers of the same sex, against all other registered speakers of the opposite sex, against k registered speaker chosen at random, against the k nearest neighbours in the registered population, etc. Whereas, in a first approximation, a population of casual registered speakers may be relatively representative of a population of cooperative registered speakers, no test protocol using casual impostors can accurately approximate the behaviour of intentional impostors. In practice, a real impostor could try to vary his voice characteristics for a fixed identity along successive trials, until he succeeds in defeating the system, gives up, or until the system blacklists the genuine user. Or he may try as many registered identities as he can with his natural voice or a disguised voice, until he succeeds, gives up, or until the police arrives!

However, most laboratory evaluations use speech databases which have usually not been recorded in a real-world situation. Therefore they do not

³¹ This concept is no doubt largely academic.

 $^{^{32}\,\}mathrm{This}$ can be the case for forensic applications.

³³Or who are not even aware that they are being recorded.

³⁴Except himself, if he is also a registered speaker.

model accurately either cooperativeness or intentional imposture, and the impostor speakers are casual impostors. A frequent practice is to use an exhaustive attempt test configuration, for which each impostor is successively tested against each registered speaker. We suggest adopting a slightly different approach. Two distinct experiments should in fact be carried out: one for which each casual impostor utterance is tested against all registered identities of the same sex, and a second one for which each casual impostor utterance is tested against all registered identities of the opposite sex. The first experiment permits estimation of the rejection ability of a system towards unacquainted intentional impostors who would know the sex of the genuine speaker, even though casual impostors are almost well-intentioned impostors. The second experiment tests whether the system is really robust to cross-sex imposture.³⁵ We will refer to these configurations as a *selec*tive attempt against all same-sex speakers and selective attempt against all cross-sex speakers respectively. In a first approximation, the proportion of successful violations does not depend on the number of registered speakers. In addition, testing each impostor utterance against its nearest neighbour in the registered population can give an indication of the system's robustness against intentional imposture. However, the result will be directly influenced by the registered speaker population size. Therefore this approach is only meaningful in the framework of a comparative evaluation on a common database. This approach can be qualified as a selective attempt towards the nearest registered neighbour. Other selective attempts are possible, such as towards speakers of the same age class, for instance.

To summarise, registered speakers should be qualified as cooperative, casual or uncooperative, whereas a distinction should be made between wellintentioned, casual or (acquainted/unacquainted) intentional impostors. Only field data can provide realistic instances of user behaviour.

Additionally, the general motivation and behaviour of the users can have an impact on the performance of a system: for instance, what are the stakes of a successful identification or verification, the benefits of an imposture, the feeling of users towards voice technology in general, etc. In evaluation, all these aspects influence the motivation of the user, and therefore the interpretation of the results.

11.3.6 Recommendations

The description of an evaluation experiment or an assessment protocol concerning a speaker classification or recognition system should explicitly report on the following items:

1. Speech quality

³⁵This step is really necessary. It can happen that a system is more robust to same-sex imposture than to cross-sex imposture, in particular if the pseudo-impostor bundle of a given speaker is only composed of speakers of the same sex (Reynolds 1994).

- 424 Part III: Spoken language system assessment
 - the general speech acquisition and transmission characteristics, in particular the signal bandwidth, the nature of the noise, the signal-tonoise ratio (SNR) when measurable, the characteristics of the transmission line, ...
 - the speech quality factors that *remain constant* across training sessions and test sessions, in particular concerning the environment, the microphone, the channel, ...
- 2. Temporal drift
 - the way the speech material is *split* between training and test material in correlation with the chronological order of its recording; for each speaker, test material should always be posterior to the latest training material.
 - the average *number of training sessions* that are necessary to register one new speaker; in practice, the number of distinct convocations he has to answer.
 - the average *registration timespan* elapsed between the first training session and the last training session, for one new speaker.
 - the average *number of test sessions* per registered user, taken into account in the evaluation.
 - the average operation timespan elapsed between the first test session and the last test session, per registered user, during the evaluation.
- 3. Speech quantity and variety
 - the average speech quantity per training session which is used per speaker for one training session, and if relevant, the average percentage of effective training speech quantity, i.e. the proportion of training speech which is actually used to build the registered speaker models.
 - the average speech quantity per test session which is used per speaker for one test session, and if relevant, the average percentage of effective speech quantity per test session, i.e. the proportion of test speech which is actually used to identify or verify the speaker, in test mode.
 - the qualitative description or characterisation of training and test *linguistic content*.
- 4. Speaker population size and typology
 - for speaker identification, the registered speaker population size, i.e. the number of registered users; this figure can also be reported for speaker verification experiments, but only as an indication of the statistical validity of the results.
 - the proportion of male and female registered speakers as well as any other relevant characteristics of the typology of registered speakers, when known; in particular, concerning the age, the dialectal origin, whether they are native or non-native speakers, etc. In parallel, any geographical, physiological, psychological or sociological feature that would be common to the registered population members (or to a

majority of them) should be identified and reported.

- for speaker verification (and open-set identification), the origin of pseudo-impostors, i.e. whether they are chosen among the registered speakers or among an external pseudo-impostor population; in the latter case, the number of external pseudo-impostors, the proportion of male and female speakers, the population typology, the speech quantity per session (per pseudo-impostor) and the number of training sessions (per pseudo-impostor) should be reported.
- for speaker verification (and open-set identification), the origin of test impostors, i.e. whether they are chosen among the registered speakers (but claiming a false identity), among the pseudo-impostors, or among an external test impostor population. The last approach is by far the most realistic. However, when it is not feasible, an impostor utterance should never be tested against a registered speaker whose bundle of pseudo-impostors contains the test impostor.
- when an external test impostor population is used, the number of external test impostors, the proportion of male and female speakers, the population typology, and in particular how their profile differs from the registered population and from the pseudo-impostor population should be described. The speech quantity per session (per test impostor) should be reported, as an indication of the statistical confidence of the evaluation results.
- 5. Speaker purpose and other human factors
 - the general purpose of the system, and whether the evaluation data are adequate for this purpose.
 - the *intention of registered speakers*, i.e. whether, in the test mode, they are cooperative speakers, uncooperative speakers, or if they behave as casual registered speakers.
 - the intention of impostors, i.e. whether they are well-intentioned impostors, casual impostors or intentional impostors. In the case of intentional impostors, the amount of knowledge they have on the true speaker should be specified, in particular whether they are acquainted by voice with the genuine speaker, and whether they are or not provided with the password, for text-dependent systems.
 - the impostor test configuration, i.e. what is the simulated (or real-life) strategy of an impostor in choosing to claim such and such an identity. For instance an exhaustive attempt, if each impostor tries each registered identity, or a selective attempt, if a certain criterion guides the impostor's choice, this criterion being stipulated. With laboratory recordings of casual speakers, we recommend the same-sex selective attempt configuration and the cross-sex selective attempt towards the nearest registered speaker, especially for comparative evaluation of two systems on a same database.
 - the stakes of the system, i.e. what are the sources of motivation for registered speakers to be recognised (or not recognised), and those of

an impostor to be accepted (or rejected).

11.3.7 Example

Here is what could be the description of an experimental protocol of an evaluation. 36

"The following protocol was designed to estimate the performance of a speaker verification system for the protection of personal portable telephones. The principle of the targeted security system is a personal-password text-dependent speaker verification system. Before a user can place a call on his portable phone, he is asked to utter his identity, i.e. his name and surname. The compatibility between the speaker and the authorised owner is checked locally, and in case of acceptance, the speaker is allowed to dial his number.

To simulate this application, the following experimental protocol was set up. A group of portable phone owners were provided with a (slightly modified) miniature tape recorder (the size of a dictaphone), and were asked to record their name and surname before they placed a call on their phone, except if they had already done so during the previous three hours. To make sure that some users would not record all their utterances consecutively, a temporisation was implemented in the tape recorder, so that a time interval of three hours had to be respected between two activations of the record function. In return for a user's participation, his subscription to the portable phone service was paid for, for six months. In practice, the six month subscription was refunded to a user when he brought back a recorded tape containing 100 recordings. This number corresponds approximately to one session every other day over six months. In reality, the average time after which a tape was returned was 4.2 months.

Once a tape and a tape recorder were returned, the tape's content was digitised at a sampling frequency of 16 kHz, and the data were segmented automatically (a beep had been internally recorded on the tape each time the "stop" button was pressed). The speech material was not verified exhaustively, but a speech activity detector was used to discard utterances that were composed of silence only. On the average, 97% of the utterances were kept. Silent signal portions lasting longer than 0.2 seconds were removed automatically. The typical bandwidth of the tape recorder's microphone is 150–6000 Hz, which is within the

 $^{^{36}\,\}mathrm{The}$ experiment reported here is hypothetical.

tape's bandwidth. All tapes were of the same trademark, and their noise level was judged negligible. Despite the fact that, for a given speaker, the microphone and the tape characteristics remained constant for all recording sessions, the data collection protocol can be considered as realistic for the targeted application.

The first five recordings for each speaker were used as training material, whereas the remaining ones were used as test material (92, on the average). The average registration timespan was estimated to be $(5/97) \times 4.2$ months ≈ 6.5 days, which may be an overestimate of the actual timespan, as users probably recorded their voice more often at the beginning of the experiment. Accordingly, the average operation timespan was considered to cover approximately four months.

An initial population of 188 persons agreed to take part in the experiment, but 19 of them never returned the recording device, either because they lost it, or because they lost interest in the experiment. Additionally, seven tape recorders and three tapes deteriorated during the six month timespan. Altogether, only 159 different speakers were thus taken into consideration as registered speakers, among which 92 were male speakers (i.e. 58%). All of them were adults over 18. Nothing else about their profile was studied, but they are likely to correspond to a relatively affluent fraction of the population, since they can afford a portable phone.

In this database, a speaker utters his name and surname in 0.8 seconds on the average, but this figure varies significantly from one person to another. The linguistic content of the speech material cannot be specified other than exhaustively.

For impostor modelling, we used all recordings corresponding to the registration phase for all speakers, which we pooled together to form a speaker-independent text-independent model. We then derived an impostor model for each registered speaker as the representation of the user's training pronunciations according to the speaker-independent model. In other words, all registered speakers were part of the pseudo-impostor bundle of a given speaker, including this very speaker.

Six professional imitators (4 male, 2 female) were then asked to simulate acquainted intentional test impostors. For each registered speaker of the same sex, they were provided with the tape recorder of the genuine user, and could listen as much as they wanted to the training material of this user. Then they were asked to produce five imitated utterances of the speaker

saying his name and surname. These imitations were recorded on the user's tape recorder, at the end of the user's tape. Given the experimental protocol, it was not possible to provide the imitators with any feedback concerning their success or failure to break the system. Altogether, each male imitator recorded approximately $5 \times 92 = 460$ impostor tests against registered male speakers, and each female speaker produced about $5 \times 67 = 335$ impostor tests against registered female speakers. All imitators were paid for their work. The imitated speech followed the same processing as the genuine one.

For the evaluation of system performance, each authentic test utterance was tried with the genuine identity $(159 \times 92 = 14628$ authentic trials), and each imitated utterance was tried against the targeted identity $(4 \times 460 + 2 \times 335 = 2510$ impostor trials)."

We leave to the reader the pleasure of tracking the unavoidable experimental biases that remain in this imaginary experimental protocol, and how they could be circumvented.

11.4 Scoring procedures

The design and implementation of an experimental evaluation protocol is usually intended to provide an estimation of a system performance. In addition to the overall figures of error rates, other relevant quantities can be derived from the analysis of the test results, for diagnostic purposes. In this section, we detail what pieces of information can be derived from evaluation results, and how to score them.

11.4.1 Notation

11.4.1.1 Registered speaker population

We will denote as "X" a population of m registered speakers:

$$X = \{X_1 \dots X_m\} = \{X_i\}_{1 \le i \le m}$$
(11.1)

For scoring purposes, we will consider a set of test utterances. We will use the term *genuine test utterances* for those which correspond to registered speakers, and the term *impostor test utterances* for those which are treated as belonging to impostor speakers. Note that the same speech segment can be used as a genuine utterance and as an impostor utterance, in some test configurations. Therefore, different notations can correspond to the same speech utterance. Each registered speaker X_i is supposed to have produced c_i genuine test utterances, the set of which will be denoted as " x_i ":

$$x_i = \{x_i^1 \dots x_i^{c_i}\} = \{x_i^k\}_{1 \le k \le c_i}$$
(11.2)

where superscript k denotes the k^{th} test utterance of speaker X_i .

In the rest of this chapter, we will denote as c the total number of genuine test utterances, and as p_i the proportion of utterances belonging to speaker X_i in the test set, that is:

$$c = \sum_{i=1}^{m} c_i \quad \text{and} \quad p_i = \frac{c_i}{c} \tag{11.3}$$

With the convention:

$$u_i^* = 1$$
 if $c_i \neq 0$ and $u_i^* = 0$ otherwise (11.4)

we will denote:

$$m^* = \sum_{i=1}^m u_i^* \tag{11.5}$$

Integer m^* is the number of registered speakers for which there is at least one genuine test utterance.

Finally, we will denote as M the set of *male* registered speakers, as F the set of *female* registered speakers and as m_M^* and m_F^* the respective number of male and female registered speakers for which there is at least one genuine test utterance.

11.4.1.2 Test impostor population

In the most general case, the whole set of impostor test utterances can be divided in subsets corresponding to one among n impostors $\{Y_j\}_{1 \le j \le n}$ using the system with a claimed identity X_i . Hence the general notation

$$y_{j\bar{\imath}} = \{y_{j\bar{\imath}}^1 \dots y_{j\bar{\imath}}^{d_{j\bar{\imath}}}\} = \{y_{j\bar{\imath}}^k\}_{1 \le k \le d_{j\bar{\imath}}}$$
(11.6)

to denote the set of $d_{j\bar{i}}$ impostor test utterances produced by impostor Y_j claiming he is X_i .

Similarly to genuine test utterances, we will denote as d the total number of impostor utterances, and as $q_{j\bar{i}}$ the proportion of impostor tests by impostor Y_j against registered speaker X_i , that is:

$$d = \sum_{j=1}^{n} \sum_{i=1}^{m} d_{j\bar{\imath}} \text{ and } q_{j\bar{\imath}} = \frac{d_{j\bar{\imath}}}{d}$$
(11.7)

We will also write:

$$d_{\bar{\imath}} = \sum_{j=1}^{n} d_{j\bar{\imath}} , \quad \tilde{d}_{j} = \sum_{i=1}^{m} d_{j\bar{\imath}} \quad \text{and} \quad q_{\bar{\imath}} = \frac{d_{\bar{\imath}}}{d} , \quad \tilde{q}_{j} = \frac{\tilde{d}_{j}}{d}$$
(11.8)

When the identity of impostors is not a relevant factor (or is unknown), subscript j can be dropped, and y_i^k denotes the k^{th} impostor attempt against registered speaker X_i $(1 \le k \le d_i)$. Conversely, in open-set speaker identification, impostors do not claim a particular identity; they just try to be identified as one of the registered speakers, whoever this speaker may be. In this case, subscript \tilde{i} can be dropped, and y_j^k denotes the k^{th} impostor attempt by impostor Y_j $(1 \le k \le \tilde{d}_j)$. If, moreover, the impostor's identity does not matter, subscript j can also be dropped, and y^k simply denotes the k^{th} impostor attempt $(1 \le k \le d)$.

With the conventions:

$$\begin{array}{lll} v_{j\bar{\imath}}^*=1 & \text{if} & d_{j\bar{\imath}}\neq 0 & \text{and} & v_{j\bar{\imath}}^*=0 & \text{otherwise} \\ v_{i}^*=1 & \text{if} & d_{\bar{\imath}}\neq 0 & \text{and} & v_{\bar{\imath}}^*=0 & \text{otherwise} \\ \tilde{v}_{i}^*=1 & \text{if} & \tilde{d}_{i}\neq 0 & \text{and} & \tilde{v}_{i}^*=0 & \text{otherwise} \end{array}$$

we will write:

$$n_{\bar{i}}^* = \sum_{j=1}^n v_{j\bar{i}}^* \qquad \tilde{m}_j^* = \sum_{i=1}^m v_{j\bar{i}}^* \qquad n^* = \sum_{j=1}^n \tilde{v}_j^* \qquad \tilde{m}^* = \sum_{i=1}^m v_{\bar{i}}^* \quad (11.9)$$

and

$$(mn)^* = \sum_{i=1}^m \sum_{j=1}^n v_{j\bar{\imath}}^*$$
(11.10)

Integer n_i^* is the number of impostors for which there is at least one test utterance against registered speaker X_i , integer \tilde{m}_j^* represents the number of registered speakers against which there is at least one test utterance from impostor Y_j , n^* is the number of impostors from which there is at least one impostor test utterance, \tilde{m}^* is the number of registered speakers against which there is at least one impostor test utterance, and $(mn)^*$ denotes the total number of couples (X_i, Y_j) for which there is at least one impostor test utterance (from Y_j against X_i).

Finally, we will denote as \tilde{M} the set of *male* impostor speakers, as \tilde{F} the set of *female* impostor speakers, and

- $\tilde{m}_{M}^{*} \colon$ the number of male registered speakers against which there is at least one impostor test utterance
- $\tilde{m}_{F}^{*}:$ the number of female registered speakers against which there is at least one impostor test utterance

- $(m_M n)^*$: the number of couples $(X_i \in M, Y_j)$ for which there is at least one impostor test utterance (from Y_j against X_i , X_i being a male speaker)
- $(m_F n)^*$: the number of couples $(X_i \in F, Y_j)$ for which there is at least one impostor test utterance (from Y_j against X_i , X_i being a female speaker)

11.4.2 Closed-set identification

A closed-set identification system can be viewed as a function which assigns, to any test utterance z, an estimated speaker index $\hat{i}(z)$, corresponding to the identified speaker $X_{\hat{i}(z)}$ in the set of registered speakers.

In closed-set identification, all test utterances belong to one of the registered speakers. Therefore, a *misclassification error* occurs for test utterance number k produced by speaker X_i when:

$$\hat{i}(x_i^k) \neq i \text{ or equivalently } \delta[\hat{i}(x_i^k), i] = 0$$
(11.11)

where δ denotes the Kronecker function, which is 1 if the two arguments are the same and 0 otherwise.

11.4.2.1 Misclassification rates

The most natural figure that indicates the performance of a speaker identification system is the relative number of times the system fails in correctly identifying an applicant speaker; in other words, how often a test utterance will be assigned an erroneous identity. Whereas it is straightforward to calculate a performance figure on a speaker-by-speaker basis, care should be taken when deriving a global score.

With our notation, and assuming that $c_i \neq 0$, we define the *misclassification* rate for speaker X_i as:

$$\gamma_i = 1 - \frac{1}{c_i} \sum_{k=1}^{c_i} \delta\left[\hat{i}(x_i^k), i\right]$$
(11.12)

If we denote as $\mathcal{P}(\hat{X}_{j\neq i} | X_i)$ the probability that the system under test identifies another speaker (with index j) than the actual speaker X_i , the quantity γ_i provides an estimate of this probability, whereas $1 - \gamma_i$ provides an estimate of $\mathcal{P}(\hat{X}_i | X_i)$. However, it is preferable to report error scores rather than success scores, and performance improvements should be measured as relative error rate reduction.³⁷ If $c_i = 0$, γ_i is undefined but $p_i \gamma_i = 0$.

 $^{^{37}}$ In fact, $log_{10} \gamma_i$ would be the most relevant figure, especially for small error rates, but the intuitive meaning of the logarithm of an error rate is less immediate. This remark also holds for mistrust rates.

We suggest using the term *dependable speaker* to qualify a registered speaker with a low misclassification rate,³⁸ and the term *unreliable speaker* for a speaker with a high misclassification rate.³⁹

From speaker-by-speaker figures, the *average misclassification rate* can be derived as:

$$\bar{\gamma} = \frac{1}{m^*} \sum_{\substack{i=1\\c_i \neq 0}}^m \gamma_i \tag{11.13}$$

and by computing separately:

$$\bar{\gamma}_{M} = \frac{1}{m_{M}^{*}} \sum_{\substack{i=1\\X_{i} \in M\\c_{i} \neq 0}}^{m} \gamma_{i} \text{ and } \bar{\gamma}_{F} = \frac{1}{m_{F}^{*}} \sum_{\substack{i=1\\X_{i} \in F\\c_{i} \neq 0}}^{m} \gamma_{i}$$
(11.14)

the gender-balanced misclassification rate can be obtained as:

$$\bar{\gamma}_{MF} = \frac{1}{2} \left(\bar{\gamma}_M + \bar{\gamma}_F \right) \tag{11.15}$$

The previous scores are different from the *test set misclassification rate*, calculated as:

$$\gamma = 1 - \frac{1}{c} \sum_{i=1}^{m} \sum_{k=1}^{c_i} \delta\left[\hat{i}(x_i^k), i\right] = \sum_{i=1}^{m} p_i \gamma_i$$
(11.16)

Scores $\bar{\gamma}$ and γ are formally identical if and only if c_i does not depend on *i*, i.e. when the test set contains an identical number $c_0 = c/m$ test utterances for each speaker. As it is usually observed that speaker recognition performances may vary with the speaker's gender, the comparison of $\bar{\gamma}$ and $\bar{\gamma}_{MF}$ can show significant differences, if the registered population is not gender-balanced. Therefore, we believe that an accurate description of the identification performance requires the three numbers $\bar{\gamma}_{MF}$, $\bar{\gamma}$, and γ to be provided.

11.4.2.2 Mistrust rates

Taking another point of view, performance scores can be designed to measure how reliable the decision of the system is when it has assigned a given identity; in other words, to provide an estimate of the probability

 $^{^{38}\}mathrm{Sometimes}$ familiarly called a sheep.

³⁹Sometimes familiarly called *a goat*.

 $\mathcal{P}(X_{j\neq i} | \hat{X}_i)$, i.e. the probability that the speaker is *not* really X_i when the system under test has output X_i as the most likely identity. To define the *mistrust rate*, we have to introduce the following notation:

$$\hat{c}_i = \sum_{k=1}^c \, \delta\left[\hat{i}(x^k), i\right] \quad \text{and} \quad \hat{p}_i = \frac{\hat{c}_i}{c}$$
(11.17)

$$\hat{m}^* = \sum_{i=1}^m \hat{m}^*_i$$
 where $\hat{m}^*_i = 1$ if $\hat{c}_i \neq 0$ and $\hat{m}^*_i = 0$ otherwise

By definition, \hat{c}_i and \hat{p}_i are respectively the number and proportion of test utterances identified as X_i over the whole test set, while \hat{m}^* is the number of registered speakers whose identity was assigned at least once to a test utterance.

The *mistrust rate* for speaker X_i can then be computed (for $\hat{c}_i \neq 0$) as:

$$\hat{\gamma}_i = 1 - \frac{1}{\hat{c}_i} \sum_{k=1}^{c_i} \delta\left[\hat{i}(x_i^k), i\right] = 1 - (1 - \gamma_i) \frac{c_i}{\hat{c}_i}$$
(11.18)

Here again, if $\hat{c}_i = 0$, $\hat{\gamma}_i$ is undefined, but $\hat{p}_i \hat{\gamma}_i = 0$.

We suggest that the term *resistant speaker* could be used to qualify a registered speaker with a low mistrust rate,⁴⁰ and the term *vulnerable speaker* for a speaker with a high mistrust rate.⁴¹

From this speaker-by-speaker score, the $average\ mistrust\ rate\ can\ be\ derived$ as:

$$\bar{\hat{\gamma}} = \frac{1}{\hat{m}^*} \sum_{\substack{i=1\\\hat{c}_i \neq 0}}^m \hat{\gamma}_i$$
(11.19)

and by computing separately:⁴²

$$\bar{\hat{\gamma}}_{M} = \frac{1}{\hat{m}_{M}^{*}} \sum_{\substack{i=1\\X_{i} \in M\\\hat{c}_{i} \neq 0}}^{m} \hat{\gamma}_{i} \text{ and } \bar{\hat{\gamma}}_{F} = \frac{1}{\hat{m}_{F}^{*}} \sum_{\substack{i=1\\X_{i} \in F\\\hat{c}_{i} \neq 0}}^{m} \hat{\gamma}_{i}$$
(11.20)

 $^{^{40}\,\}mathrm{The}$ familiar term of ram could be used to extend the bovine analogy.

 $^{^{41}\,\}mathrm{The}$ familiar term of lamb seems appropriate here.

 $^{^{42}}$ With \hat{m}_M^* and \hat{m}_F^* being respectively the number of male and female speakers whose identity was assigned at least one to a test utterance.

the gender-balanced mistrust rate is defined as:

$$\bar{\hat{\gamma}}_{MF} = \frac{1}{2} \left(\bar{\hat{\gamma}}_M + \bar{\hat{\gamma}}_F \right) \tag{11.21}$$

By noticing now that:

$$\sum_{i=1}^{m} \hat{p}_i \,\hat{\gamma}_i = \sum_{i=1}^{m} p_i \,\gamma_i \tag{11.22}$$

there appears to be no need to define a test set mistrust rate $\hat{\gamma}$. In other words: the test set mistrust rate is equal to the test set misclassification rate.⁴³

From a practical point of view, misclassification rates and mistrust rates can be obtained by the exact same scoring programs, operating successively on the confusion matrix and on its transpose.

11.4.2.3 Confidence ranks

Most speaker identification systems use a similarity measure between a test utterance and all training patterns to decide, by a nearest neighbour decision rule, which is the identity of the test speaker. In this case, for a test utterance x, an ordered list of registered speakers can be produced:

$$\hat{L}(x) = [X_{l_1(x)} \dots X_{l_m(x)}] = [X_{l_j(x)}]_{1 \le l_j(x) \le m}$$
(11.23)

where, for all index $j,\;X_{l_j(x)}$ is judged closer to the test utterance than $X_{l_{j+1}(x)}$ is.

The *identification rank* of the genuine speaker of utterance x_i^k can then be expressed as:

$$\rho(x_i^k) = \sum_{j=1}^m j \cdot \delta[l_j(x_i^k), i] \text{ or equivalently}
\rho(x_i^k) = \underset{j=1}{\operatorname{argmax}} \delta[l_j(x_i^k), i]$$
(11.24)

In other words, ρ is the position at which the correct speaker appears in the ordered list of neighbours of the test utterance. Note that a correct identification of x_i^k corresponds to $\rho(x_i^k) = 1$.

Under the assumption that $c_i \neq 0$, let us now denote:

$$\eta_{(r)}(x_i^k) = \sum_{j=1}^r \delta[l_j(x_i^k), i] \text{ and } \eta_{i(r)} = \frac{1}{c_i} \sum_{k=1}^{c_i} \eta_{(r)}(x_i^k)$$
(11.25)

⁴³This property can be interesting to debug a scoring program.

i.e.:
$$\eta_{(r)}(x_i^k) = 1$$
 if $\rho(x_i^k) \leq r$ and 0 otherwise

The α % confidence rank for speaker X_i , which we will denote here as $\rho_{i(\alpha \%)}$ can then be defined as the smallest integer number for which α % of the test utterances belonging to speaker X_i are part of the ρ_i nearest neighbours in the ordered list of candidates. Hence the formulation:

$$\rho_{i(\alpha\%)} = \min_{r} \{ r \mid \eta_{i(r)} \ge \alpha\% \}$$
(11.26)

Then, the average $\alpha \%$ confidence rank can be computed as the average of $\rho_{i(\alpha \%)}$ over all registered speakers (for which $c_i \neq 0$):

$$\bar{\rho}_{(\alpha\,\%)} = \frac{1}{m^*} \sum_{\substack{i=1\\c_i \neq 0}}^{m} \rho_{i\,(\alpha\,\%)}$$
(11.27)

Though a gender-balanced confidence rank could be defined analogously to gender-balanced misclassification and mistrust rates, the relevance of this figure is not clear.

If finally we denote:

$$\eta_{(r)} = \sum_{i=1}^{m} \sum_{k=1}^{c_i} \eta_{(r)} \left(x_i^k \right) = \sum_{i=1}^{m} p_i \eta_{i(r)}$$
(11.28)

the test set $\alpha \%$ confidence rank is defined as:

$$\rho_{(\alpha \%)} = \min_{r} \{ r \mid \eta_{(r)} \ge \alpha \% \}$$
(11.29)

11.4.2.4 Comments

Average scores, gender-balanced scores and test set scores all fall under the same formalism. If we denote as π_i a certain quantity which we will call the *relative representativity* of speaker X_i and which satisfies $\sum \pi_i = 1$, and if we now consider the linear combination:

$$\gamma_{\pi} = \sum_{i=1}^{m} \pi_i \gamma_i \tag{11.30}$$

It is clear that:

$$\gamma_{\pi} = \bar{\gamma} \quad \text{for } \pi_{i} = \frac{1}{m^{*}}$$

$$\gamma_{\pi} = \bar{\gamma}_{MF} \quad \text{for } \pi_{i} = \frac{1}{2m_{M}^{*}} \text{ if } X_{i} \in M \quad \text{and} \quad \pi_{i} = \frac{1}{2m_{F}^{*}} \text{ if } X_{i} \in F$$

$$\gamma_{\pi} = \gamma \quad \text{for } \pi_{i} = p_{i}$$

Therefore $\bar{\gamma}$, $\bar{\gamma}_{MF}$ and γ correspond to different estimates of a global score, under various assumptions on the relative representativity of each speaker.⁴⁴ For average scores, it is assumed that each speaker is equally representative, irrespectively of its sex group, but if the test population is strongly unbalanced this hypothesis may not be relevant (unless there is a reason for it). For gender-balanced scores, each test speaker is supposed to be representative of its sex group, and each sex group is supposed to be equiprobable. Test set scores make the assumption that each test speaker has a representativity which is proportional to its number of test utterances, which is certainly not always a meaningful hypothesis.

Test set scores can therefore be used as an overall performance measure if the test data represent accurately the profile and behaviour of the user population, both in terms of population composition and individual frequency of use. If only the composition of the test set population is representative of the general user population, average scores allow neutralisation of the possible discrepancies in number of utterances per speaker. If finally the composition of the test set speaker population is not representative, gender-balanced scores provide a general purpose estimate.

If there is a way to estimate separately the relative representativity for each test registered speaker, a *representative mi sclassification rate* can be computed as in equation (11.30). Conversely, some techniques such as those used in public opinion polls can be resorted to in order to select a representative test population when setting up an evaluation experiment.

11.4.2.5 Example

Tables 11.2 and 11.3 give examples of misclassification rates, mistrust rates and confidence ranks. How ever, it must be kept in mind that the number of tests used to design these examples is too small to guarantee any statistical validity of the figures.

11.4.3 Verification

A verification system can be viewed as a function which assigns, to a test utterance z and a claimed identity \tilde{i} , a boolean value $\hat{v}_{\tilde{i}}(z)$, which is equal to 1 if the utterance is accepted, and 0 if it is rejected.

Two types of error can then occur. Either a genuine speaker is rejected, or an impostor is accepted. Hence, a *false rejection* corresponds to:

$$\hat{v}_{\bar{\imath}}\left(x_i^k\right) = 0 \tag{11.31}$$

and a *false acceptance* happens if:

$$\hat{v}_{\bar{i}} \left(y_{j\bar{i}}^k \right) = 1$$
 (11.32)

 $^{^{44}}$ This remark is also true for mistrust rates, with the estimates of π_i being obtained from the output of the identification system.

m = 5	X_1	X_2	X_3	X_4	X_5	\hat{c}_i	$\hat{\gamma}_i$	
	Male speakers		Female speakers					
$\hat{X_1}$	4	_	_	_	_	4	0%	$\hat{m}^* = 5$
$\hat{X_2}$	-	0	1	-	1	2	100%	$ar{\hat{\gamma}}=30~\%$
$\hat{X_3}$	-	_	1	-	-	1	0%	$\hat{m}_M^* = 3 \qquad \hat{m}_F^* = 2$
$\hat{X_4}$	4	1	-	5	-	10	50%	${ar {ar \gamma}}_M pprox 33\% {ar {ar \gamma}}_F pprox 25\%$
$\hat{X_5}$	-	-	I		3	3	0%	$\bar{\hat{\gamma}}_{MF}\approx 29\%$
c_i	8	1	2	5	4	c = 20		
γ_i	50%	100%	50%	0%	25%			
	$m^* = 5$							
	$\bar{\gamma} = 45 \%$							
	$\hat{m}_M^* = 3 \qquad \qquad \hat{m}_F^* = 2$							
	$ \bar{\gamma}_M \approx 67 \% \qquad \bar{\gamma}_F \approx 13 \% $						$\gamma\equiv\hat{\gamma}$	$=\frac{7}{20}=35\%$
	$\bar{\gamma}_{MF} \approx 40 \%$							

Table 11.2: Example of misclassification & mistrust rate scoring froma confusion matrix in speaker identification.

Here X_4 would be a *dependable* speaker, X_2 an *unreliable* speaker, X_1 , X_3 and X_5 would appear as *resistant* speakers and X_2 (him again!) would seem to be a *vulnerable* speaker.

Test	Reco	gnised	speak	eratr	ank j	id. rank	conf. rank	conf. rank
utterance	$j \equiv 1$	$j \equiv 2$	$j \equiv 3$	$j \equiv 4$	$j \equiv 5$	$\rho(x_i^{n})$	at 50%	at 95%
x_1^1	X_1	X_2	X_4	X_3	X_5	1		
x_{1}^{2}	X_1	X_4	X_2	X_5	X_3	1		
x_1^3	X_4	X_1	X_3	X_2	X_5	2		
x_{1}^{4}	X_4	X_1	X_2	X_5	X_3	2		
x_{1}^{5}	X_1	X_4	X_5	X_2	X_3	1		
x_1^6	X_4	X_1	X_2	X_3	X_5	2		
x_1^7	X_1	X_4	X_3	X_2	X_5	1		
x_{1}^{8}	X_4	X_1	X_3	X_2	X_5	2		
$\eta_{1(j)}$	50%	100%	100%	100%	100%	\rightarrow	$ \rho_{1 (50\%)} = 1 $	$\rho_{1~(95\%)}=2$
r^1	X 4	Xr	Xa	X_{2}	X_1	3		
$\eta_{2(i)}^{w_2}$	0%	0%	100%	100%	100%	\rightarrow	$\rho_{2(50\%)} = 3$	$\rho_{2(95\%)} = 3$
. – (5)							(/)	(, 0)
x_{3}^{1}	X_2	X_1	X_3	X_4	X_5	3		
x_{3}^{2}	X_3	X2	X_1	X_4	X_{5}	1	o 1	
$\eta_{3(j)}$	50%	50%	100%	100%	100%	\rightarrow	$ ho_{3\ (50\%)}=1$	$ ho_{3\ (95\%)}=3$
x_4^1	X_4	X_1	X_3	X_2	X_5	1		
x_{4}^{2}	X_4	X_2	X_5	X_3	X_1	1		
x_{4}^{3}	X_4	X_1	X_5	X_3	X_2	1		
x_4^4	X_4	X_2	X_3	X_5	X_1	1		
x_{4}^{5}	X_4	X_2	X_5	X_3	X_1	1		
$\eta_{4(j)}$	100%	100%	100%	100%	100%	\rightarrow	$ \rho_{4\ (50\%)} = 1 $	$ \rho_{4\ (95\%)} = 1 $
$x^{\frac{1}{2}}$	X_5	X_{2}	X_{2}	X_4	X_1	1		
$x_{\rm E}^2$	X_5	X_3	X_1	X_4	X_2	1		
x_{E}^{3}	X_2	X_4	X_3	X_1	$\begin{bmatrix} x_5 \end{bmatrix}$	5		
x_{5}^{4}	$\overline{X_5}$	X_2	X_1	X_4	X_3	1		
$\eta_{5(j)}$	75%	75%	75%	75%	100%	\rightarrow	$ ho_{5\ (50\%)}=1$	$ ho_{5\ (95\%)}=5$
							· ↓ ·	· ↓ ·
							$\rho_{(50\%)} = 1.2$	$ ho_{(95\%)}=2.8$
$n_{(i)}$	65%	85%	95%	95%	100%		$a_{(50\%)} = 1$	$\theta(0 = 0) = 3$
· ()	55,0	50,0	20,0	50,0	20070	\rightarrow	r (00%)	r (99%)

Table 11.3: Example of confidence rank scoring in closed-set speaker identification

Only for speaker X_2 is the right speaker in first position ($\rho_{i (50\%)} = 1$ less than half the time except for i = 2). On the test set, a list of the 3 best candidates would have to be kept to ensure that the right speaker is in the list 95% of the time ($\rho_{test}(95\%) = 3$).

In the rest of this section we will denote the events as follows:

- \hat{V} : the system accepts the applicant speaker
- \hat{V} : the system rejects the applicant speaker
- X: the applicant speaker is a genuine speaker
- Y: the applicant speaker is an impostor $(Y \equiv \overline{X})$
- *i* or *j*: the identity of the applicant speaker is X_i or Y_j
- $\tilde{\imath}$: the claimed identity is i

We first address aspects of *static evaluation*, that is, what meaningful figures can be computed to measure the performance of a system over which the experimentator has absolutely no control. Then, after discussing the role of *decision thresholds*, we review several approaches that allow a *dynamic evaluation* of the system to be obtained, i.e. in a relatively threshold-independent manner.

11.4.3.1 False rejection rates

If $c_i \neq 0$, the *false rejection rate* for speaker X_i is defined quite naturally as:

$$\alpha_i = 1 - \frac{1}{c_i} \sum_{k=1}^{c_i} \hat{v}_{\bar{i}}(x_i^k)$$
(11.33)

Rate α_i provides an estimate of $\mathcal{P}(\hat{V} | X, i)$, i.e. the probability that the system makes a diagnostic of rejection, given that the applicant speaker is the authorised speaker X_i (claiming his own identity). If $c_i = 0$, α_i is undefined but $p_i \alpha_i = 0$.

As for closed-set identification, the terms *dependable* speaker and *unreliable* speaker can be used to qualify speakers with a low or (respectively) high false rejection rate.

From speaker based figures, the *average false rejection rate* can be obtained as:

$$\bar{\alpha} = \frac{1}{m^*} \sum_{\substack{i=1\\c_i \neq 0}}^{m} \alpha_i$$
(11.34)

while the gender-balanced false rejection rate is:

$$\bar{\alpha}_{MF} = \frac{1}{2} \left(\bar{\alpha}_M + \bar{\alpha}_F \right) \tag{11.35}$$
where :

$$\bar{\alpha}_{M} = \frac{1}{m_{M}^{*}} \sum_{\substack{i=1\\X_{i} \in M\\c_{i} \neq 0}}^{m} \alpha_{i} \text{ and } \bar{\alpha}_{F} = \frac{1}{m_{F}^{*}} \sum_{\substack{i=1\\X_{i} \in F\\c_{i} \neq 0}}^{m} \alpha_{i} \quad (11.36)$$

The test set false rejection rate is calculated as:

$$\alpha = 1 - \frac{1}{c} \sum_{i=1}^{m} \sum_{k=1}^{c_i} \hat{v}_i(x_i^k) = \sum_{i=1}^{m} p_i \alpha_i$$
(11.37)

Rates $\bar{\alpha}$, $\bar{\alpha}_{MF}$ and α provide three different estimates of $\mathcal{P}(\ddot{V} | X)$. Rate α is influenced by the test set distribution of genuine attempts which may only be artefactual.

11.4.3.2 False acceptance rates and imposture rates

As opposed to false rejection, there are several ways to score false acceptance, depending on whether it is the vulnerability of registered speakers which is considered or the skills of impostors. Moreover, the way to evaluate false acceptance rates and imposture rates depends on whether the identity of each impostor is known or not.

KNOWN IMPOSTORS

If the impostor identities are known, the false acceptance rate in favour of impostor Y_j against registered speaker X_i can be defined, for $d_{j\bar{i}} \neq 0$, as:

$$\beta_{j\bar{\imath}} = \frac{1}{d_{j\bar{\imath}}} \sum_{k=1}^{d_{j\bar{\imath}}} \hat{v}_{\bar{\imath}} \left(y_{j\bar{\imath}}^k \right)$$
(11.38)

Here, $\beta_{j\bar{\imath}}$ can be viewed as an estimate of $\mathcal{P}(\hat{V} | Y, j, \hat{\imath})$, i.e. the probability that the system makes a diagnostic of acceptance, given that the applicant speaker is the impostor Y_j claiming identity X_i .

Then, the average false acceptance rate against speaker X_i can be obtained (if $n_i^* \neq 0$) by averaging the false acceptances over all impostors:

$$\beta_{\bar{i}} = \frac{1}{n_{\bar{i}}^{*}} \sum_{\substack{j=1\\d_{j\bar{i}} \neq 0}}^{n} \beta_{j\bar{i}}$$
(11.39)

and similarly the average imposture rate in favour of impostor Y_j can be calculated (for $\tilde{m}_j^* \neq 0$) as:

$$\tilde{\beta}_j = \frac{1}{\tilde{m}_j^*} \sum_{\substack{i=1\\d_{j\bar{\imath}} \neq 0}}^m \beta_{j\bar{\imath}}$$
(11.40)

Assessment of speaker verification systems 441

Rates β_i and $\tilde{\beta}_j$ provide (respectively) estimates of $\mathcal{P}(\hat{V} | Y, \hat{\imath})$ and $\mathcal{P}(\hat{V} | Y, j)$ under the assumption that all impostors and all claimed identities are equiprobable. The number $\beta_{\tilde{\imath}}$ indicates the false acceptance rate obtained on average by each impostor in claiming identity X_i , while $\tilde{\beta}_j$ indicates the success rate of impostor Y_j in claiming an identity averaged over each claimed identity. A registered speaker can be more or less resistant (low $\beta_{\tilde{\imath}}$) or vulnerable (high $\beta_{\tilde{\imath}}$), whereas impostors with a high $\tilde{\beta}_j$ can be viewed as skilled impostors,⁴⁵ as opposed to poor impostors⁴⁶ for those with a low $\tilde{\beta}_j$. The average false acceptance rate which is equal to the average imposture rate is obtained as:

$$\bar{\beta} = \frac{1}{(mn)^*} \sum_{\substack{(i,j) = (1,1) \\ d_{j\bar{\imath}} \neq 0}}^{(m,n)} \beta_{j\bar{\imath}} = \bar{\tilde{\beta}}$$
(11.41)

i.e. as the average of the false acceptances over all couples (X_i, Y_j) ,⁴⁷ which provides an estimate of $\mathcal{P}(\hat{V} | Y)$ under the assumption that all couples (X_i, Y_j) are equally likely.

Here, separate estimates of the average false acceptance rate on the male and female registered populations can be obtained as:

$$\bar{\beta}_{M} = \frac{1}{(m_{M}n)^{*}} \sum_{\substack{(i, j) = (1, 1) \\ X_{i} \in M \\ d_{ji} \neq 0}}^{(m, n)} \beta_{ji}$$

and $\bar{\beta}_{F} = \frac{1}{(m_{F}n)^{*}} \sum_{\substack{(i, j) = (1, 1) \\ X_{i} \in F \\ d_{ji} \neq 0}}^{(m, n)} \beta_{ji}$ (11.43)

⁴⁵ Wolves, in our animal analogy.

⁴⁶ Badgers seems particularly appropriate!

⁴⁷ Another way of estimating the average false acceptance rate and the average imposture rate could consist in computing:

$$\bar{\beta}^{(1)} = \frac{1}{\tilde{m}^*} \sum_{\substack{i=1\\ d_{\tilde{i}} \neq 0}}^{m} \beta_{\tilde{i}} \quad \text{and} \quad \tilde{\bar{\beta}}^{(1)} = \frac{1}{n^*} \sum_{\substack{j=1\\ \tilde{d}_{j} \neq 0}}^{n} \tilde{\beta}_{j} \quad (11.42)$$

but these estimates have the drawback that, in general, $\bar{\beta}^{(1)} \neq \bar{\tilde{\beta}}^{(1)}$.

and a gender-balanced false acceptance rate is provided by:

$$\bar{\beta}_{MF} = \frac{1}{2} \left(\bar{\beta}_M + \bar{\beta}_F \right) \tag{11.44}$$

The question could be raised of whether it is desirable to compute a score which would provide an estimation of the false acceptance rate for a genderbalanced impostor population. We propose not to go that far, as it would clearly lead to duplication of scoring figures, but the influence of impostors' gender could be partly neutralised by the experimental design:

- If the impostor population is composed of *acquainted intentional* impostors, the issue of impostor's gender balancing can be considered as relatively marginal, even though the impostor of a given sex may be more skilled in imitating a same-sex person than somebody of the opposite sex.
- If the impostor population is composed of *casual* impostors, we propose to restrict systematically the impostor utterance test set to same-sex trials. However, as we mentioned above, it is safer to check in an independent experiment whether the system is really robust to cross-sex casual impostors.

It may also be interesting to calculate imposture rates regardless of the claimed identities. In this case, we define the *imposture rate* in favour of impostor Y_j regardless of the claimed identity as:

$$\tilde{\beta}'_{j} = \frac{1}{\tilde{d}_{j}} \sum_{k=1}^{d_{j}} \hat{v}_{i} \left(y_{j}^{k} \right)$$
(11.45)

and the average imposture rate regardless of the claimed identity as:

$$\bar{\tilde{\beta}}' = \frac{1}{n^*} \sum_{\substack{j=1\\ \tilde{d}_j \neq 0}}^n \tilde{\beta}'_j$$
(11.46)

However, $\beta_{j\bar{\imath}}$, $\beta_{\bar{\imath}}$, $\tilde{\beta}_{j}$, $\tilde{\beta}_{j}$, $\bar{\beta}$, $\bar{\beta}_{MF}$ and $\bar{\beta}'$ cannot be evaluated when the identities of impostors are not known. In this case false acceptance rates and imposture rates can be calculated under the assumption that all impostor test utterances are produced by distinct impostors.

UNKNOWN IMPOSTORS

The false acceptance rate against speaker X_i assuming distinct impostors can be obtained (if $d_i^* \neq 0$) as:

$$\beta_{\bar{i}}' = \frac{1}{d_{\bar{i}}} \sum_{k=1}^{d_{\bar{i}}} \hat{v}_{\bar{i}} \left(y_{\bar{i}}^k \right) \tag{11.47}$$

and the *average false acceptance rate assuming distinct impostors* is defined as:

$$\bar{\beta}' = \frac{1}{\tilde{m}^*} \sum_{\substack{i=1\\d_i \neq 0}}^m \beta_i'$$
(11.48)

Here again, separate estimates of the average false acceptance rate assuming distinct impostors, on the male and female registered populations can be obtained as:

$$\bar{\beta}'_{M} = \frac{1}{\tilde{m}^{*}_{M}} \sum_{\substack{i=1\\X_{i} \in M\\d_{\tilde{i}} \neq 0}}^{m} \beta'_{i} \text{ and } \bar{\beta}'_{F} = \frac{1}{\tilde{m}^{*}_{F}} \sum_{\substack{i=1\\X_{i} \in F\\d_{\tilde{i}} \neq 0}}^{m} \beta'_{i} \quad (11.49)$$

with the gender-balanced false acceptance rate assuming distinct impostors being:

$$\bar{\beta}'_{MF} = \frac{1}{2} \left(\bar{\beta}'_{M} + \bar{\beta}'_{F} \right) \tag{11.50}$$

Rate β'_i provides a speaker-dependent estimate of \mathcal{P} $(\hat{V} \mid Y, \tilde{\imath})$ assuming distinct impostors. Rate $\bar{\beta}'$ can be viewed as an estimate of \mathcal{P} $(\hat{V} \mid Y)$ under the assumptions of distinct impostors and that all claimed identities are equally likely while $\bar{\beta}'_{MF}$ can be understood as another estimate of \mathcal{P} $(\hat{V} \mid Y)$ under the assumptions of distinct impostors, that attempts against male speakers and against female speakers are equiprobable, and that within a gender class all claimed identities are equally likely.

Test set scores

If finally false acceptances are scored globally, regardless of the impostor identity nor of the claimed identity, we obtain the *test set false acceptance* rate which is identical to the *test set imposture rate*:

$$\beta = \frac{1}{d} \sum_{k=1}^{d} \hat{v}(y^{k}) = \frac{1}{d} \sum_{i=1}^{m} \sum_{k=1}^{d_{\tilde{i}}} \hat{v}_{\tilde{i}}(y^{k}_{\tilde{i}}) = \tilde{\beta}$$
(11.51)

Here, $\beta \equiv \tilde{\beta}$ provides a test set estimate of $\mathcal{P}(\hat{V} | Y)$ which is biased by the composition of the registered population and a possible uneveness of the number of impostor trials for each speaker. Note the relations

$$\beta_{\bar{i}}' = \sum_{j=1}^{n} \frac{q_{j\bar{i}}}{q_{\bar{i}}} \beta_{j\bar{i}} , \qquad \tilde{\beta}_{j}' = \sum_{i=1}^{m} \frac{q_{j\bar{i}}}{\tilde{q}_{j}} \beta_{j\bar{i}} \quad \text{and} \beta = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{j\bar{i}} \beta_{j\bar{i}} = \sum_{i=1}^{m} q_{\bar{i}} \beta_{i}' = \sum_{j=1}^{n} q_{j} \tilde{\beta}_{j}'$$
(11.52)

SUMMARY

For scoring false acceptance rates, we believe that, beside β , it is necessary to report on $\overline{\beta}$ and $\overline{\beta}_{MF}$ (when impostors are known) or $\overline{\beta}'$ and $\overline{\beta}'_{MF}$ (when they are not known), as the score β may be significantly influenced by the test data distribution. The other scores described in this section are mainly useful for diagnostic analysis.

11.4.3.3 Relative unreliability, vulnerability and imitation ability

It can also be of major interest to estimate the contribution of a given registered speaker X_i to the overall false rejection rate, which can be denoted as $\mathcal{P}(i | \hat{V}, X)$, i.e. the probability that the identity of the speaker was *i* given that a (false) rejection diagnostic was made on a genuine speaker (claiming his own identity).

We can thus define the average relative unreliability for speaker X_i as:

$$\bar{\mu}_i = \frac{\alpha_i}{\sum_i \alpha_i} = \frac{1}{m^*} \frac{\alpha_i}{\bar{\alpha}}$$
(11.53)

or his test set relative unreliability:

$$\mu_i = \frac{p_i \,\alpha_i}{\sum_i \, p_i \,\alpha_i} = p_i \,\frac{\alpha_i}{\alpha} \tag{11.54}$$

By construction:

$$\sum_{i} \bar{\mu}_{i} = \sum_{i} \mu_{i} = 1 \tag{11.55}$$

From a different angle, the relative vulnerability for a given registered speaker X_i (i.e. $\mathcal{P}(\tilde{i} | \hat{V}, Y)$) can be measured as his contribution to the false acceptance rate.

Thus, the average relative vulnerability for speaker X_i can be defined as:

$$\bar{\nu}_{\bar{\imath}} = \frac{\sum_{j} \beta_{j\bar{\imath}}}{\sum_{i} \sum_{j} \beta_{j\bar{\imath}}} = \frac{n_{\bar{\imath}}^*}{(mn)^*} \frac{\beta_{\bar{\imath}}}{\bar{\beta}}$$
(11.56)

his relative vulnerability assuming distinct impostors, as:

$$\bar{\nu}_{i}' = \frac{\beta_{i}'}{\sum_{i} \beta_{i}'} = \frac{1}{\tilde{m}^{*}} \frac{\beta_{i}'}{\bar{\beta}'}$$
(11.57)

and his test set relative vulnerability as:

$$\nu_{\bar{\imath}} = \frac{\sum_{j} q_{j\bar{\imath}} \beta_{j\bar{\imath}}}{\sum_{i} \sum_{j} q_{j\bar{\imath}} \beta_{j\bar{\imath}}} = q_{\bar{\imath}} \frac{\beta'_{\bar{\imath}}}{\beta}$$
(11.58)

Here:

$$\sum_{i} \bar{\nu}_{\bar{i}} = \sum_{i} \bar{\nu}'_{\bar{i}} = \sum_{i} \nu_{\bar{i}} = 1$$
(11.59)

Finally, by considering the relative success of impostor Y_j , i.e. $\mathcal{P}(j | \hat{V}, Y)$, we define in a dual way, as above, the *average imitation ability* of impostor Y_j :

$$\bar{\tilde{\nu}}_{j} = \frac{\sum_{i} \beta_{j\bar{\imath}}}{\sum_{j} \sum_{i} \beta_{j\bar{\imath}}} = \frac{\tilde{m}_{j}^{*}}{(mn)^{*}} \frac{\tilde{\beta}_{j}}{\bar{\beta}}$$
(11.60)

his imitation ability regardless of the claimed identity:

$$\bar{\tilde{\nu}}'_{j} = \frac{\tilde{\beta}'_{j}}{\sum_{j} \tilde{\beta}'_{j}} = \frac{1}{n^{*}} \frac{\tilde{\beta}'_{j}}{\bar{\tilde{\beta}}'}$$
(11.61)

and his test set relative imitation ability:

$$\tilde{\nu}_{j} = \frac{\sum_{i} q_{j\bar{\imath}} \beta_{j\bar{\imath}}}{\sum_{j} \sum_{i} q_{j\bar{\imath}} \beta_{j\bar{\imath}}} = \tilde{q}_{j} \frac{\tilde{\beta}'_{j}}{\beta}$$
(11.62)

Naturally:

$$\sum_{j} \bar{\tilde{\nu}}_{j} = \sum_{j} \bar{\tilde{\nu}}'_{j} = \sum_{j} \tilde{\nu}_{j} = 1$$
(11.63)

The relative unreliability and vulnerability can also be calculated relatively to the male/female population.

11.4.3.4 Comments

As for misclassification rates, the gender-balanced, average and test set false rejection rates as well as the gender-balanced and average false acceptance rates assuming distinct impostors and the test set false acceptance rate correspond to different estimates of a global score, under various assumptions on the relative representativity of each genuine test speaker. The discussion of Section 11.4.2.4 can be readily generalised.

For what concerns gender-balanced and average false acceptance rates with known impostors, a relative representativity $\kappa_{j\bar{i}}$ can be defined for each couple of registered speaker and impostor (X_i, Y_j) (with $\sum \sum \kappa_{j\bar{i}} = 1$), and if we write:

$$\beta_{\kappa} = \sum_{i=1}^{m} \sum_{j=1}^{n} \kappa_{j\bar{\imath}} \beta_{j\bar{\imath}}$$
(11.64)

we have:

$$\beta_{\kappa} = \bar{\beta} \quad \text{for} \quad \kappa_{j\bar{\imath}} = \frac{1}{(mn)^*}$$

$$\beta_{\kappa} = \bar{\beta}_{MF} \quad \text{for} \quad \kappa_{j\bar{\imath}} = \frac{1}{2(m_M n)^*} \text{ if } X_i \in M \quad \text{and} \quad \kappa_{j\bar{\imath}} = \frac{1}{2(m_F n)^*} \text{ if } X_i \in F$$

In the case of casual impostors, choosing a selective attempt configuration towards same-sex speakers is equivalent to the assumption:

 $\begin{bmatrix} X_i \in M \text{ and } Y_j \in \tilde{F} \end{bmatrix} \text{ or } \begin{bmatrix} X_i \in F \text{ and } Y_j \in \tilde{M} \end{bmatrix} \implies \kappa_{j\tilde{\imath}} = 0 \tag{11.65}$

i.e. that the representativity of a cross-sex attempt is zero.

Studies allowing better definition of the representativity of impostor attempts against registered speakers would be of great help to increase the relevance of evaluation scores.

11.4.3.5 Example

Tables 11.4, 11.5 and 11.6 give examples of false acceptance rates, false rejection rates, and imposture rates, as well as unreliability, vulnerability and imitation ability. As for the closed-set identification examples, the number of tests used to design these examples is too small to guarantee any statistical validity.

11.4.3.6 Expected benefit

From now on, we will denote as α and β the false rejection and acceptance rates, whichever exact estimate is really chosen. Estimates of the following quantities are required:

- p, the probability that an applicant speaker is a genuine speaker,
- $B_X \ge 0$, the benefit of a true acceptance,
- $B_Y \ge 0$, the benefit of a true rejection,
- $C_X \ge 0$, the cost of a false rejection,
- $C_Y \ge 0$, the cost of a false acceptance.

	m = 3			
	X_1	X_2	X_3	
	Ma	ale	\mathbf{Female}	
Ŷ	6	2	4	
$\hat{ar{V}}$	3	0	3	
c_i	9	2	7	
c	18			
m^*	3			
$lpha_i$	$\frac{3}{9} = \frac{1}{3}$	0	$\frac{3}{7}$	
$\bar{\alpha}$	$\frac{1}{3}\left[\frac{3}{9} + \frac{0}{2} + \frac{3}{7}\right] = \frac{16}{63} \approx 25 \%$			
α	$\frac{3+0+3}{18} = \frac{1}{3} \approx 33\%$			
$m^*_{\scriptscriptstyle M} \mid m^*_{\scriptscriptstyle F}$	2 1			
$\bar{\alpha}_{\scriptscriptstyle M} \bar{\alpha}_{\scriptscriptstyle F}$	$\frac{1}{2}\left[\frac{1}{3}+0\right] = \frac{1}{6}$			
$\bar{\alpha}_{MF}$	$\frac{1}{2} \left[\frac{1}{6} + \frac{3}{7} \right] = \frac{25}{84} \approx 30 \%$			
$\bar{\mu}_i$	$\frac{7}{16} \approx 0.44$	0	$\frac{9}{16} \approx 0.56$	
μ_i	$\frac{1}{2} = 0.50$	0	$\frac{1}{2} = 0.50$	

Table 11.4: Genuine attempts

Out of 18 genuine attempts, 6 false rejections are observed, hence the *test set* false acceptance rate $\alpha = \frac{1}{3} \approx 33 \%$. Nevertheless, the 3 false rejections out of 9 trials for X_1 do not have the same impact on the average false rejection rate $\bar{\alpha} \approx 25 \%$ as the 3 false rejections out of 7 trials for X_3 . In fact, while X_2 seems to be the most reliable speaker, X_3 appears more unreliable than X_1 on the average, as, for what concerns relative unreliability scores, $\bar{\mu}_3 > \bar{\mu}_1$.

	m = 3						
	$Y_{\tilde{1}}$ $Y_{\tilde{2}}$			$Y_{ ilde{3}}$			
	Against Male			Against Female			
	<i>n</i> =	= 2	<i>n</i> =	= 2	<i>n</i> =	n = 2	
	$Y_{1 \tilde{1}}$	$Y_{2\tilde{1}}$	$Y_{1\tilde{2}}$	$Y_{2\tilde{2}}$	$Y_{1\;\tilde{3}}$	$Y_{2\tilde{3}}$	
\hat{V}	2	0	5	1	-	1	
$\hat{ar{V}}$	4	2	1	2	-	4	
$d_{j\widetilde{\imath}}$	6	2	6	3	-	5	
$d_{\widetilde{\imath}}$	8		9		5		
d	22						
$n_{ ilde{\imath}}^*$	4	2	2	2	1		
$ ilde{m}^*$			ŝ	3			
$(mn)^*$	5						
$eta_{j ilde{\imath}}$	$\frac{2}{6} = \frac{1}{3}$	0	<u>5</u> 6	$\frac{1}{3}$	undef.	$\frac{1}{5}$	
$eta_{ ilde{\imath}}$	$\frac{1}{2}\left[\frac{1}{3}+\right]$	$0 = \frac{1}{6}$	$\frac{1}{2}\left[\frac{5}{6}+ \cdot\right]$	$\left[\frac{1}{3}\right] = \frac{7}{12}$	$\frac{1}{1}$ $\begin{bmatrix} 1\\5 \end{bmatrix}$	$=\frac{1}{5}$	
$\bar{\beta}$	$\frac{1}{5} \left[\frac{1}{3} + 0 + \frac{5}{6} + \frac{1}{3} + \frac{1}{5} \right] = \frac{17}{50} = 34\%$						
$\beta'_{\tilde{i}}$	$\frac{2+0}{8} = \frac{1}{4} \qquad \frac{5+1}{9} = \frac{2}{3}$			-	5		
$\bar{\beta}'$	$\frac{1}{3} \left[\frac{1}{4} + \frac{2}{3} + \frac{1}{5} \right] = \frac{67}{180} \approx 37\%$						
β	$\frac{2+0+5+1+0+1}{22} = \frac{9}{22} \approx 41\%$						
$ ilde{m}^*_M \mid ilde{m}^*_F$	2			1			
$(m_M n)^* \mid (m_F n)^*$	4 1			1			
$\bar{\beta}_M \mid \bar{\beta}_F$	$\frac{1}{4} \left[\frac{1}{3} + 0 + \frac{5}{6} + \frac{1}{3} \right] = \frac{3}{8} \qquad \qquad \frac{1}{1} \left[\frac{1}{5} \right] = \frac{1}{5}$			$=\frac{1}{5}$			
$\bar{\beta}_{MF}$	$\frac{1}{2} \left[\frac{3}{8} + \frac{1}{5} \right] = \frac{23}{80} \approx 29\%$						
$\bar{\beta}'_M \mid \bar{\beta}'_F$	$\frac{1}{2}\left[\frac{1}{4} + \frac{2}{3}\right] = \frac{11}{24}$			$\frac{1}{1} \left[\frac{1}{5} \right] = \frac{1}{5}$			
$\bar{\beta}'_{MF}$	$\frac{1}{2} \left[\frac{11}{24} + \frac{1}{5} \right] = \frac{79}{240} \approx 33\%$						
$\bar{ u}_{ ilde{\imath}}$	$\frac{10}{51} \approx 0.19$		$rac{35}{51}pprox 0.69$		$\frac{2}{17} = \frac{6}{51} \approx 0.12$		
$\bar{\nu}'_{\tilde{i}}$	$\frac{15}{67} \approx 0.22$		$rac{40}{67}pprox 0.60$		$\frac{12}{67} \approx 0.18$		
$ u_{\tilde{i}}$	$\frac{2}{9} \approx 0.22$		$\frac{2}{3} = \frac{6}{9} \approx 0.67$		$\frac{1}{9} \approx 0.11$		

Table 11.5: Impostor attempts against registered speakers

One out of three impostor trials from Y_1 against X_1 were successful while none from Y_2 were. Hence $\beta_{\bar{1}} = \frac{1}{6}$. But if the identities of impostors are not known, it can only be measured that, out of 8 impostor attempts against X_1 , 2 were successful, i.e. $\beta'_{\bar{1}} = \frac{1}{4}$. As no impostor attempt from Y_1 against X_3 was recorded, the average false acceptance rate against X_3 can only be averaged over 1 impostor. Hence $\beta_{\bar{3}} = \beta'_{\bar{3}} = \frac{1}{5}$. The 3 ways of computing false acceptance rates, namely the average false acceptance rate $\bar{\beta}$, the average false acceptance rate assuming distinct impostors $\bar{\beta}'$ and the test set false acceptance rate β provide significantly different scores, as the number of test utterances is not balanced across all possible couples (X_i, Y_j) . In this example, the relative vulnerability scores $\bar{\nu}_i$, P'_i and ν_i indicate that speaker X_3 would appear as the most resistant, while speaker X_2 would seem to be the most vulnerable.

	n = 2					
	Y_1			Y_2		
	m = 3		m = 3			
	Y _{1 Ĩ}	$Y_{1\tilde{2}}$	Y _{1 3}	Y _{2Ĩ}	$Y_{2\tilde{2}}$	Y _{2 3}
Ŷ	2	5	-	0	1	1
$\hat{ar{V}}$	4	1	-	2	2	4
$d_{j\tilde{\imath}}$	6	6	-	2	3	5
$ ilde{d}_j$		12			10	
d			2	2		
$ ilde{m}_j^*$	2				3	
$\binom{n}{(mn)^*}$				5		
(1111)			,	, I		
$eta_{j ilde{\imath}}$	$\frac{2}{6} = \frac{1}{3}$	<u>5</u> 6	undef.	0	$\frac{1}{3}$	$\frac{1}{5}$
$ ilde{eta}_j$	$\frac{1}{2}\left[\frac{1}{3} + \frac{5}{6}\right] = \frac{7}{12}$			$\frac{1}{3}\left[0 + \frac{1}{3} + \frac{1}{5}\right] = \frac{8}{45}$		
$ ilde{eta}_j'$	$\frac{2+5}{12} = \frac{7}{12}$			$\frac{0+1+1}{10} = \frac{1}{5}$		
$\bar{\tilde{eta}}$	$\frac{1}{5} \left[\frac{1}{3} + \frac{5}{6} + 0 + \frac{1}{3} + \frac{1}{5} \right] \equiv \bar{\beta} = \frac{17}{50} = 34\%$					
$\bar{\tilde{eta}}'$	$\frac{1}{2} \left[\frac{7}{12} + \frac{1}{5} \right] = \frac{47}{120} \approx 39 \%$					
$ ilde{eta}$	$\frac{2+5+0+0+1+1}{22} \equiv \beta = \frac{9}{22} \approx 41\%$					
$ar{ ilde{ u}}_j$	$\frac{35}{51} \approx 0.69$			$\frac{16}{51} \approx 0.31$		
$\bar{ ilde{ u}}_j'$	$\frac{35}{47} \approx 0.74$			$\frac{12}{47} \approx 0.26$		
$ ilde{ u}_j$	$\frac{7}{9} \approx 0.78$			$\frac{2}{9} \approx 0.22$		

Table 11.6: Impostor attempts from impostors

Out of 6 trials from impostor Y_1 against speaker X_1 , 2 of them turned out to be successful, while out of 6 other trials against X_2 , 5 lead to a (false) acceptance. As no attempts from Y_1 against X_3 were recorded, the *average imposture rate* from impostor Y_1 can be estimated as $\tilde{\beta}_1 = \frac{1}{2} \left[\frac{2}{6} + \frac{5}{6}\right] = \frac{7}{12}$. If we now ignore the actual identities of violated speakers, and we summarise globally the success of impostor Y_1 , we get $\tilde{\beta}'_1 = \frac{2+5}{12}$ which turns out to be also equal to $\frac{7}{12}$. While $\tilde{\beta} = \tilde{\beta}$ and $\tilde{\beta} = \beta$, the *average imposture rate regardless of the claimed identity* $\tilde{\beta}'$ indicates that the "average" impostor is successful almost 2 times out of 5 in his attempts. All estimates of the *relative imitation ability* $(\tilde{\nu}_j, \tilde{\nu}'_j \text{ and } \tilde{\nu}_j)$ agree that Y_1 is a much more *skilled* impostor than Y_2 who seems to be quite *poor*.

With the estimates the *expected benefit* \mathcal{B} of a verification system with false rejection rate α and false acceptance rate β can be computed as:

$$\mathcal{B} = p \cdot ((1 - \alpha) B_X - \alpha C_X) + (1 - p) \cdot ((1 - \beta) B_Y - \beta C_Y)$$
(11.66)

In particular, when $p = 1 - p = \frac{1}{2}$ and $B_X = B_Y = C_X = C_Y = K$, the equal-risk equal-cost expected benefit is:

$$\mathcal{B}_{eq} = K \left(1 - (\alpha + \beta) \right) \tag{11.67}$$

The expected benefit is usually a meaningful static evaluation figure for the potential clients of the technology. It must however be understood only as the average expected benefit for each user attempt. It does not take into account external factors such as the psychological impact of the system, its maintenance costs, etc.

11.4.3.7 Threshold setting

Speaker verification systems usually proceed in two steps. First, a matching score $s(z, X_i)$ is computed between the test utterance z and the reference model X_i corresponding to the claimed identity. Then, the value of the matching score is compared to a threshold θ_i , and a decision is taken as follows:

$$f \quad s(z, X_i) \ge \theta_i \quad \hat{v}_i(z) = 1 \quad \text{else} \quad \hat{v}_i(z) = 0 \tag{11.68}$$

In other words, verification is positive only if the match between the test utterance and the reference model (for the claimed identity) is close enough. A distinction can be made depending on whether each registered speaker has his individual threshold or whether a single threshold is used which is common to all speakers. In other words, if θ_i depends on *i*, the system uses *speaker-dependent thresholds*, whereas if θ_i does not depend on *i*, the system uses a *speaker-independent threshold*. ⁴⁸ We will denote as Θ the threshold vector $[\theta_i]_{1 \leq i \leq m}$, and as α (Θ) and β (Θ) the false rejection and acceptance rates corresponding to Θ .

The values of Θ have an inverse impact on the false rejection rate and on the false acceptance rate. Thus, with a low θ_i , fewer genuine attempts from speaker X_i will be rejected, but more impostors will be erroneously accepted as X_i . Conversely, if θ_i is increased, β_i will generally decrease, at the expense of an increasing α_i . The goal of *dynamic evaluation* is to provide a description of the system performance which is as independent as possible of the threshold values.

⁴⁸ In theory, a system using speaker-dependent thresholds should never perform worse than the same system using a speaker-independent threshold, as the latter is finally a particular case of the former. However, difficulties may arise in obtaining reliable estimates of each threshold, in the speaker-dependent approach.

The setting of thresholds is conditioned to the specification of an operating constraint which expresses the compromise that has to be reached between the two types of error. Among many possibilities, the most popular ones are:

- A specified false rejection rate α_0 . If speaker-dependent thresholds are used, the thresholds are tuned so that the false rejection rate for each speaker is equal to α_0 whereas, with speaker-independent thresholds, the constraint is only satisfied on the average.
- A specified false acceptance rate β_0 . Here also, the constraint can be satisfied for each speaker with speaker-dependent thresholds, or in the average for speaker-independent thresholds.
- The maximisation of the expected benefit. Once again, the corresponding α and β can be obtained by a speaker-by-speaker optimisation or on an average basis.
- An equal error rate (or EER) α = β = ε. In fact, this is the most popular constraint, as the equal error rate ε is seen as a simple way of summarising the overall performance of a system into a single figure. Moreover, for any threshold Θ:

if
$$\alpha(\Theta) \ge \epsilon$$
 then $\beta(\Theta) \le \epsilon$
if $\alpha(\Theta) \le \epsilon$ then $\beta(\Theta) \ge \epsilon$

In most practical applications, however, the equal error rate does not correspond to an interesting operating constraint.

Two procedures are classically used to set the thresholds: the *a priori threshold setting* procedure and the *a posteriori threshold setting* procedure. When the *a priori* threshold setting procedure is implemented, the threshold vector $\Theta = \Theta_{prio}$ is estimated from a set of *tuning data*, which can be either the training data themselves, or a new set of unseen data. Then, the false rejection and acceptance rates α (Θ_{prio}) and β (Θ_{prio}) are estimated on a disjoint test set. Naturally, there must be no intersection between the tuning data set and the test data set. Not only must the speech material of genuine attempts and impostor attempts be different between these two sets, but also the bundle of pseudo-impostors used to tune the threshold for a registered speaker should not contain any of the impostors which will be tested against this very speaker within the test set. Of course, the volume of additional speech data used for threshold setting must be counted as

training material, when reporting on the training speech quantity. When the *a posteriori* threshold setting procedure is adopted, $\Theta = \Theta_{post}$ is set on the test data themselves. In this case, the false rejection and acceptance rates α (Θ_{post}) and β (Θ_{post}) must be understood as the performance of the system with ideal thresholds. Though this procedure does not lead to a fair measure of the system performance, it can be interesting, for diagnostic evaluation, to compare α (Θ_{prio}) and β (Θ_{prio}) with α (Θ_{post}) and β (Θ_{post}).

11.4.3.8 System operating characteristic

Whichever operating constraint is chosen to tune the thresholds is only one of the infinite number of possible trade-offs, and it is generally not possible to predict, from the false rejection and false acceptance rates obtained for a particular functioning point, what would be the error rates for another functioning point. In order to be able to estimate the performance of the system under any conditions, its behaviour has to be modelled so that its performance can be characterised independently from any threshold settings.

Speaker-independent threshold

In the case of a speaker-independent threshold, the false rejection and the false acceptance rates can be written as functions of a single parameter θ , namely $\alpha = \alpha$ (θ) and $\beta = \beta$ (θ). Then, a more compact way of summarising the system's behaviour consists in expressing β directly as a function of α (or the opposite), that is:

$$\beta = f(\alpha) \tag{11.69}$$

Using terminology derived from Communication Theory, function f is sometimes called the *Receiver Operating Characteristic* and the corresponding curve \mathcal{F} the *ROC curve*. Generally, function f is monotonically decreasing and satisfies the limit conditions f(0) = 1 and f(1) = 0. Figure 11.1 depicts a typical ROC curve.



Figure 11.1: A typical ROC curve and its equal error rate

The point-by-point knowledge of function f provides a thresholdindependent description of all possible functioning conditions of the system. In particular:

- If a false rejection rate α_0 is specified, the corresponding false acceptance rate is obtained as $\beta_0 = f(\alpha_0)$. Graphically, this corresponds to the intersection of the ROC curve \mathcal{F} with the vertical straight line of equation $\alpha = \alpha_0$.
- If a false acceptance rate β_0 is specified, the corresponding false rejection rate is obtained as $\alpha_0 = f^{-1}(\beta_0)$. Graphically, this corresponds to the intersection of the ROC curve \mathcal{F} with the horizontal straight line of equation $\beta = \beta_0$.
- If the expected benefit is to be maximised, the derivation of equation (11.66) shows that:

$$\frac{\partial \mathcal{B}}{\partial \theta} = 0 \implies \frac{\partial \beta}{\partial \alpha} = -\frac{p}{1-p} \frac{B_X + C_X}{B_Y + C_Y} = \lambda \tag{11.70}$$

Graphically, the corresponding functioning point is obtained by sliding, from the origin, a straight line with slope λ , until it becomes tangent to the ROC curve. The point of contact then indicates the corresponding α and β .

• To obtain the equal error rate ϵ , the equation $\epsilon = f(\epsilon)$ has to be solved. This functioning point corresponds to the intersection of the ROC curve with the straight line of equation $\beta = \alpha$.

In practice, there are several ROC curves, depending on what type of false rejection and acceptance scores are used: 49

a gender-balanced ROC: $\bar{\beta}_{MF} = f(\bar{\alpha}_{MF})$ (or $\bar{\beta}'_{MF} = f(\bar{\alpha}_{MF})$ if impostors are unknown), an average ROC $\bar{\beta} = f(\bar{\alpha})$ (or $\bar{\beta}' = f(\bar{\alpha})$ if impostors are unknown),

a test set ROC $\beta = f(\alpha)$.

However, exhaustively keeping a whole ROC curve lacks conciseness, and it is classically felt desirable to condense system performance into a single figure. Traditionally, the EER is chosen for this purpose, In this case, there is a distinct equal error rate for each ROC curve, which can be denoted by $\bar{\varepsilon}_{MF}$, $\bar{\varepsilon}$ and ε , respectively.

Speaker-dependent thresholds

In the case of speaker-dependent thresholds, the false rejection and the false acceptance rates for each speaker X_i depend on a different parameter θ_i . Therefore, each speaker has his own ROC curve:

$$\beta_i = f_i\left(\alpha_i\right) \tag{11.71}$$

 $^{^{49}\,\}mathrm{Any}$ other combination would not make much sense.

In this case, there is no simple way of deriving an "average" ROC curve that would represent the general behaviour of the system. Current practice consists in characterising each individual ROC curve by its equal error rate ξ_i , and in summarising the performance of the system by the *average equal error rate* $\bar{\xi}$ computed as:

$$\bar{\xi} = \frac{1}{m} \sum_{i=1}^{m} \xi_i$$
(11.72)

Note here that a gender-balanced equal error rate $\bar{\xi}_{MF}$ can be defined as:

$$\bar{\xi}_{MF} = \frac{1}{2} \left(\bar{\xi}_M + \bar{\xi}_F \right) = \frac{1}{2} \left(\frac{1}{m_M} \sum_{i \in M}^m \xi_i + \frac{1}{m_F} \sum_{i \in F}^m \xi_i \right)$$
(11.73)

and a *test set equal error rate* as:

$$\xi = \sum_{i=1}^{m} p_i \,\xi_i \tag{11.74}$$

Though we used the same terminology for denominating equal error rates with speaker-dependent and speaker-independent thresholds, it must be stressed that the scores are not comparable. Therefore it should always be specified in which framework they are computed.

11.4.3.9 System characteristic modelling

Equal error rates can be interpreted as a very local property of the ROC curve. In fact, as the ROC curve usually has its concavity turned in the direction of the axis $\alpha = \beta$, the EER gives an idea of how close the ROC curve is to the axes. However, this is a very incomplete picture of the general system performance level, as it is virtually impossible to predict the performance of the system under a significantly different operating condition.

Recent work by Oglesby (1994) has addressed the question of how to encapsulate the entire system characteristic into a single number. Oglesby's suggestions, which we will develop now, consist in finding a simple 1-parameter model which describes as accurately as possible the ROC curve over most of its definition domain. If the approximation is good enough, reasonable error rate estimates for any functioning point can be derived. As in the last section, we will first discuss the case of a system with a speaker-independent threshold, and then extend the approach to speaker-dependent thresholds. For modelling the relation between α and β , the simplest approach is to

assume a linear operating characteristic, i.e. a relation between α and β of the kind:

$$\alpha + \beta \approx 2\hat{\epsilon}_{lin} \quad \Longleftrightarrow \quad \beta \approx f(\alpha) = 2\hat{\epsilon}_{lin} - \alpha \tag{11.75}$$

where $\hat{\epsilon}_{lin}$ is a constant which can be understood as the *linear-model EER*.⁵⁰ However, typical ROC curves do not have a linear shape at all, and this model is too poor to be effective over a large domain.

A second possibility is to assume that the ROC curve has the approximate shape of the positive branch of a hyperbola, which supposes the relation:

$$\alpha \cdot \beta \approx \hat{\epsilon}_{hyp}^2 \iff \beta \approx \hat{f}(\alpha) = \frac{\hat{\epsilon}_{hyp}^2}{\alpha}$$
(11.76)

Here $\hat{\epsilon}_{hyp}$ is another constant which can be interpreted as the *hyperbolic*model *EER*. The hyperbolic model is equivalent to a linear model in the log-error domain. It usually fits the ROC curve much better.⁵¹ However, it has the drawback of not fulfilling the limit conditions, as $\hat{f}(0) = +\infty$ and $\hat{f}(1) \neq 0$.

A third possibility, proposed by Oglesby, is to use the following model:

$$\alpha^{-\frac{1}{\log_2 \hat{\epsilon}_{ogl}}} + \beta^{-\frac{1}{\log_2 \hat{\epsilon}_{ogl}}} \approx 1 \quad \Longleftrightarrow \quad \beta \approx \hat{f}(\alpha) = \left(1 - \alpha^{-\frac{1}{\log_2 \hat{\epsilon}_{ogl}}}\right)^{-\log_2 \hat{\epsilon}_{ogl}} (11.77)$$

where $\hat{\epsilon}_{ogl}$ will be referred to as $Oglesby's\ model\ EER$. Oglesby reports a good fit of the model with experimental data, and underlines the fact that $\hat{f}\left(0\right)=1$ and $\hat{f}\left(1\right)=0.$

The parametric approach is certainly a very relevant way to give a broader system characterisation. Nevertheless, several issues remain questionable.

First, it is clear that none of the models proposed above account for a possible skewness of the ROC curve. As Oglesby notes it, to address skewed characteristics would require introducing an additional variable, which would give rise to a second, non-intuitive, figure.

A second question is what criterion $C(\hat{f})$ should be minimised to fit the model curve $\hat{\mathcal{F}}$ to the true ROC curve \mathcal{F} . If we denote as $[\alpha_1, \alpha_2]$ the *optimisation domain* on which the best fit is to be found, the most natural criterion would be to minimise the mean square error between $\hat{\mathcal{F}}$ and \mathcal{F} over the interval $[\alpha_1, \alpha_2]$. However, an absolute error difference does not have the same meaning when α changes order of magnitude, and an alternative could be to minimise the mean square error between the curves in a log–log representation.

A third and most crucial question is how the unavoidable deviations between the model and the actual ROC curve should be quantified and reported.

 $^{^{50}}$ This approach is to be linked to a common practice which consists in summarising the performance of a speaker verification system as the arithmetic mean of the false rejection rate and the false acceptance rate obtained for a particular threshold.

 $^{^{51}}$ Hence, the geometric mean is usually a better way to average a false rejection rate and a false acceptance rate for a given threshold.

Here is a possible answer to these questions. Though the approach that we are going to present has not been extensively tested so far, we believe that it is worth exploring it in the near future, as it may prove useful to summarise concisely the performance of a speaker verification system, in a relatively meaningful and exploitable manner.

The solution proposed starts by fixing an *accuracy* ω % for the ROC curve modelling, say for instance 5 %. Then, if we define:

$$\begin{aligned} \phi_{\omega}^{+}\left(\alpha\right) &= \left(1+\omega\right)f\left(\alpha\right) & \phi_{\omega}^{-}\left(\alpha\right) = \left(1-\omega\right)f\left(\alpha\right) \\ \psi_{\omega}^{+}\left(\alpha\right) &= f\left(\frac{\alpha}{1+\omega}\right) & \psi_{\omega}^{-}\left(\alpha\right) = f\left(\frac{\alpha}{1-\omega}\right) \end{aligned}$$

the following properties are obvious:

then
$$\frac{\left|\hat{f}\left(\alpha\right) - f\left(\alpha\right)\right|}{\left|f\left(\alpha\right)\right|} = \frac{\left|\hat{f}\left(\alpha\right) - \beta\right|}{\beta} \le \omega$$
(11.78)

then
$$\frac{\left|\hat{f}^{-1}\left(\beta\right) - f^{-1}\left(\beta\right)\right|}{f^{-1}\left(\beta\right)} = \frac{\left|\hat{f}^{-1}\left(\beta\right) - \alpha\right|}{\alpha} \le \omega$$
(11.79)

(11.80)

Hence, when both constraints are satisfied, both relative differences between the modelled and exact false rejection and acceptance rates are below ω %. Then, a model of the ROC curve must be chosen, for instance Oglesby's model. However, if another model fits the curve better, it can alternatively be used, but it preferably should depend on a single parameter, and the link between the value of this parameter and the model equal error rate should be specified.

For a given parameter $\hat{\epsilon}$, the lower and upper bound of the ω %-accuracy false rejection rate validity domain, $\alpha_{\omega}^{inf}(\hat{\epsilon})$ and $\alpha_{\omega}^{sup}(\hat{\epsilon})$ are obtained by decreasing (or increasing) α , starting from the initial value $\alpha_{init} = \epsilon$, until one of the two constraints of equations (11.78) and (11.80) is no more satisfied. This process can be repeated for several values of $\hat{\epsilon}$ varying for instance in small steps within the interval [$\epsilon (1 - \omega), \epsilon (1 + \omega)$]. Finally, the value of $\hat{\epsilon}$ corresponding to the wider validity domain can be chosen as the system performance measure, in the validity domain of the approximation. Note that $\hat{\epsilon}$ does not need to be inside the validity domain for its value to be meaningful. If the validity domain turns out to be too small, then the process could be repeated after having set the accuracy ω to a higher value. Another possibility could be to give several model equal error rates, corresponding to several adjacent validity domains (with a same accuracy ω), i.e. a piecewise representation of the ROC curve.

A first advantage of the parametric description is that it allows prediction of the behaviour of a speaker verification system for a more or less extended set of operating conditions. It could then be possible to give clear answers to a potential client of the technology, as long as this client is able to specify his constraints. The second advantage is that the model EER is a number which relates well to the conventional EER. Therefore the new description would not require that the scientific community totally changes its point of view in apprehending the performance of a speaker verification system.⁵² The main drawback of the proposed approach is that it lacks experimental validation for the time being. Therefore, we suggest adopting it as an experimental evaluation methodology, until it has been proven efficient.

In dealing with a system using speaker-dependent thresholds, we are brought back to the difficulty of averaging ROC curve models across speakers. The ROC curve for each speaker X_i can be summarised by a model equal error rate $\hat{\xi}_i$ and a ω %-accuracy false acceptance rate validity domain $[\alpha_{i\omega}^{inf}, \alpha_{i\omega}^{sup}]$. Lacking a more relevant solution, we suggest characterising the average system performance by averaging across speakers the model EER, and the bounds of the validity interval. Thus the global system performance could be given as an average model EER:

$$\bar{\hat{\xi}} = \frac{1}{m} \sum_{i=1}^{m} \hat{\xi}_i$$
(11.81)

and an average ω %-accuracy false acceptance rate validity domain:

$$\left[\bar{\alpha}_{\omega}^{inf}, \bar{\alpha}_{\omega}^{sup}\right] \tag{11.82}$$

where:

$$\bar{\alpha}_{\omega}^{inf} = \frac{1}{m} \sum_{i=1}^{m} \alpha_{i\,\omega}^{inf} \quad \text{and} \quad \bar{\alpha}_{\omega}^{sup} = \frac{1}{m} \sum_{i=1}^{m} \alpha_{i\,\omega}^{sup}$$
(11.83)

The same approach can be implemented, with different weights, to compute a gender-balanced model EER $\overline{\hat{\xi}}_{MF}$ and a test set model EER $\hat{\xi}$, and the corresponding validity domains.

Another possibility would be to fix a speaker-independent validity domain $[\alpha^{inf}, \alpha^{sup}]$ for each ROC curve, and then compute the individual accuracy ω_i . Then, to obtain a global⁵³ score, all ω_i could be averaged (using

⁵²In fact, a conventional EER ε is the model EER $\hat{\varepsilon}$ for any model, but with a 0%-validity domain of $[\varepsilon, \varepsilon]$!

⁵³That is gender-balanced, average or test set.

weights depending on the type of estimate), and the performance would be a global model equal error rate together with a false acceptance rate domain $[\alpha^{inf}, \alpha^{sup}]$ common to all speakers, but at an average accuracy.

11.4.3.10 Example

For example, consider a verification system with a speaker-independent threshold that has a gender-balanced Oglesby's equal error rate of 4.7 % with a 5%-accuracy false rejection rate validity domain of [0.5%, 25.0%]. Here, the ROC curve under consideration is $\bar{\beta}_{MN} = f(\bar{\alpha}_{MN})$. We will denote now $\bar{\alpha}_{MN} = a$ and $\bar{\beta}_{MN} = b$, for simplicity reasons.

For any false rejection rate a satisfying $0.005 \leq a \leq 0.250$, the difference between the actual false acceptance rate b and the estimated false acceptance rate $\hat{f}(a)$ predicted by Oglesby's model with parameter $\hat{\varepsilon}_{MF} = 0.047$ satisfies $|b - \hat{f}(a)| \leq 0.05 b$. It can then be computed (using equation (11.77)) that the 5%-accuracy false *acceptance* rate validity domain is [0.3%, 20.6%], and it is guaranteed that, for any value of b in this interval, the difference between the actual false rejection rate a and the estimated false rejection rate $\hat{f}^{-1}(b)$ (predicted by Oglesby's model with EER 0.047) satisfies $|a - \hat{f}^{-1}(b)| \leq 0.05 a$. In particular, the exact (gender-balanced) EER of the system, $\bar{\varepsilon}_{MF}$, is equal to 0.047, at a 5% relative accuracy.

11.4.4 Open-set identification

An open-set identification system can be viewed as a function which assigns to any test utterance z an estimated speaker index $\hat{w}(z) \in \{1...m\}$, corresponding to the identified speaker $X_{\hat{w}(z)}$ in the set of registered speakers, or outputs 0 if the applicant speaker is considered as an impostor. In open-set identification, three types of error can be distinguished:

• a misclassification error occurs for a genuine test utterance x_i^k when:

$$\hat{w}\left(x_{i}^{k}\right) \neq 0 \quad \text{and} \quad \hat{w}\left(x_{i}^{k}\right) \neq i$$

$$(11.84)$$

• a false rejection occurs for a genuine test utterance x_i^k when:

$$\hat{w}(x_i^k) = 0$$
 (11.85)

• and a *false acceptance* occurs if, for an impostor test utterance y_i^k :

$$\hat{w}\left(y_{i}^{k}\right) \neq 0 \tag{11.86}$$

Here, two points of view can be adopted.

Either a misclassification error is considered as a false acceptance (while a correct identification is treated as a true acceptance). In this case, openset identification can be scored in the same way as verification, namely by evaluating a false rejection rate α and a false acceptance rate β . The concept

of ROC curve can be extended to this family of systems, and in particular, an equal error rate ϵ can be computed. However, the false acceptance rate β is now bounded by a value $\gamma^c \leq 1$ when the threshold θ tends to 0, γ^c being the closed-set misclassification rate of the system, i.e. the performance that the open-set identification system would provide if it was functioning in a closed-set mode. Therefore, a parametric approach for dynamic evaluation would require a specific class of ROC curve models (at least with two parameters). Moreover, merging classification errors with false acceptances may not be appropriate if the two types of error are not equally harmful.

An alternative solution is to keep distinct the three types of error, and measure them by three rates α , β and γ . The ROC curve is now a curve in a three-dimensional space, with equation $\varphi(\alpha, \beta, \gamma) = 0$. The two extremities of this curve are the points with coordinates ($\alpha = 1, \beta = 0, \gamma = 0$) and ($\alpha = 0, \beta = 1, \gamma = \gamma^c$). The ROC curve can be projected as $\beta = f(\alpha)$ and $\gamma = g(\beta)$. The first projection is a monotonically decreasing curve such as f(0) = 1 and f(1) = 0, whereas the second projection is also monotonically decreasing, and satisfies g(0) = 0 and $g(1) = \gamma^c$. A minimal description of the curve of φ could then be the equal error rate ϵ of function f and the closed-set identification score γ^c of function g. Parametric models of φ with two degrees of freedom could be thought of, but to our knowledge, this remains an unexplored research topic.

Among both possibilities, we believe that the second one is to be preferred, though it is slightly more complex.

11.4.5 Recommendations

These recommendations indicate how the performance of a speaker recognition system should be scored.

- 1. For closed-set identification
 - Beside the test set misclassification rate (γ) , report on average misclassification and mistrust rates $(\bar{\gamma} \text{ and } \bar{\hat{\gamma}})$, and provide also genderbalanced rates $(\bar{\gamma}_{MF} \text{ and } \bar{\hat{\gamma}}_{MF})$ if the test population is composed of male and female speakers.
 - As the number of registered speakers is a crucial factor of performance, it is essential to indicate the number of speakers in the registered speaker population. Mention also the proportion of male and female speakers for information.
 - For statistical validity information, indicate the number and male/female proportion of speakers in the test population and the average number of test utterances per test speaker.
- 2. For verification
 - For static evaluation, beside the test set false rejection rate (α) and the test set false acceptance rate (β), provide the average false rejection rate ($\bar{\alpha}$) and the average false acceptance rate ($\bar{\beta}$ or $\bar{\beta}'$ depending

on whether the impostors' identities are known or not). Genderbalanced rates $(\bar{\alpha}_{MF} \text{ and } \bar{\beta}_{MF} \text{ or } \bar{\beta}'_{MF})$ should also be reported.

 For dynamic evaluation and a speaker-independent threshold, the system ROC curve should be obtained as:

either $\bar{\beta}_{MF} = f(\bar{\alpha}_{MF})$ (or $\bar{\beta}'_{MF} = f(\bar{\alpha}_{MF})$ if impostors are unknown),

or $\bar{\beta} = f(\bar{\alpha})$ (or $\bar{\beta}' = f(\bar{\alpha})$ if impostors are unknown), or $\beta = f(\alpha)$

Summarise a ROC curve by its traditional equal error rate (respectively $\bar{\varepsilon}_{MF}$, $\bar{\varepsilon}$ and ε). Investigate on the possibility of finding a ROC curve model, and report the model equal error rate $(\hat{\overline{\varepsilon}}_{MF}, \hat{\varepsilon} \text{ and } \hat{\varepsilon})$ and the ω %-accuracy false rejection rate validity domain $[\alpha^{inf}, \alpha^{sup}]$. Find a reasonable compromise between ω and $\alpha^{sup} - \alpha^{inf}$.

- For dynamic evaluation with speaker-dependent thresholds, compute the individual equal error rate (ξ_i) of each ROC curve $\beta_i = f(\alpha_i)$ and give the gender-balanced equal error rate $(\bar{\xi}_{MF})$, the average equal error rate $(\bar{\xi})$ and the test set equal error rate (ξ) . Investigate the possibility of fitting a common ROC curve model by adjusting individually a model equal error rate $(\hat{\xi}_i)$ for each curve. Here, either an accuracy ω is fixed and speaker-dependent validity domains are computed, or the validity domain is fixed in a speaker-independent manner and the individual accuracy ω_i is computed. Compute anyway the global model equal error rates $(\bar{\xi}_{MF}, \bar{\xi} \text{ and } \hat{\xi})$. Then give accordingly either the average validity domain for a speaker-independent accuracy or the average accuracy for a speaker-independent validity domain.
- For statistical validity information, indicate the number of registered speakers, the proportion of male and female registered speakers, the number of genuine test speakers, the proportion of male and female genuine test speakers and the average number of genuine test utterances per genuine test speaker. Give also a relevant description of the test impostor configuration and population.
- 3. For open-set identification
 - For static evaluation, score separately the false rejections (α , $\bar{\alpha}$ and $\bar{\alpha}_{MF}$), the false acceptances (β , $\bar{\beta}$ or $\bar{\beta}'$ and $\bar{\beta}_{MF}$ or $\bar{\beta}'_{MF}$) and the misclassifications (γ , $\bar{\gamma}$ and $\bar{\gamma}_{MF}$).
 - For dynamic evaluation and a speaker-independent threshold, project the three-dimensional ROC curve $\varphi(\alpha, \beta, \gamma)$ into two curves $\beta = f(\alpha)$ and $\gamma = g(\beta)$. Summarise the first one as its equal error rate ε and the second one as its extremity γ^c . Investigate the possibility of using a parametric approach.
 - For dynamic evaluation with speaker-dependent thresholds, average individual ξ_i and individual γ_i^c . Investigate the possibility of using a parametric approach.
 - As for closed-set identification and for verification, give all relevant in-

formation concerning the registered population, the genuine test population and the impostor population and test configuration.

In practice, gender-balanced average and test set scores are obtained very easily as various linear combinations of individual speaker scores.

11.5 Comparative and indirect assessment

11.5.1 Reference systems

In the previous sections, the focus was centered on performance evaluation of the algorithms with emphasis on scoring procedures. However, in most cases, absolute performance may be misleading as it depends on many aspects of the experimental conditions. This is particularly true when the evaluation is based on training and test data which may not be publicly available. In such a case, it is impossible to reproduce the test independently.

In fact, frequent scientific practice consists in comparing a new approach to a previously existing one, and in evaluating the new method in terms of (relative) improvement. This practice can be understood as a way to calibrate the difficulty of the task under test by some kind of reference technique. In this section, we propose to use common reference algorithms in order to make the calibration step even more meaningful.

Desirable properties of a reference system are its relative efficiency and robustness, its easy implementation, and its absolute reproducibility (from the algorithmic point of view). It should not require sophisticated training procedures. In particular, it should be able to operate with very limited training data (like one utterance) per speaker.

From a practical point of view, the calibration of a given database for a given task by a reference system is relatively easy to implement: for any new system embedded in a given application, the reference voice recognition system is replaced by the new one, and the differences in the overall performance of the application gives an indirect figure of merit of the new system compared to the reference one, for this given application. Commercial products can also be evaluated in parallel on the same database, but this may cause additional difficulties. In particular, the development of a specific "harness", i.e. an interface between the evaluation data and the system under test may be necessary.

In general, a speaker recognition system can be decomposed into:

- 1. a pre-processing module, which extracts acoustic parameters from the speech signal (this module includes voice activity detection),
- $2. \$ a speaker modelling module, which computes some kind of speaker model,
- 3. a scoring module, which computes a resemblance score between a test and a reference pattern,
- 4. a decision module, which outputs a diagnostic (identity assignment, ac-

ceptance, rejection, doubt, ...).

It is certainly unrealistic to specify exhaustively the four modules of a reference system, especially the first and fourth module, which may have considerable impact on the performance. We therefore restrict our proposal for a reference system to generic algorithms involved in the second and third module. However, if the system under test is a "black box", and if it is not possible to isolate the pre-processing module, some arbitrary choice has to be made for what concerns the pre-processing module of the reference system.

In practice, it is necessary to distinguish between two reference systems: one for text-dependent applications and one for text-independent applications. For text-dependent and text-prompted applications, a baseline system based on Dynamic Time Warping (DTW) offers a number of advantages:

- Training can be done with only one utterance per speaker, but multiple references can be used if more training data are available.
- The local constraints fully specify the algorithms; in particular, the performance is not dependent on any initialisation condition or training algorithm.
- The distance measure between patterns can be approximately interpreted as the logarithm of a likelihood.

However, DTW is very sensitive to end-point detection, and relatively heavy in computation when the number of references is high. Nevertheless, we believe that this family of algorithms offers a good compromise to obtain a reference performance on most databases for text-dependent and textprompted applications. The choice of DTW as a reference system was already proposed in the context of the SAM-A project (SAM-A 1993; Homayounpour et al. 1993).

For text-independent applications, a sphericity measure was also proposed in SAM-A as a reference method. In fact, the sphericity measure is one possibility among a large family of measures based on second-order statistics (SOS). SOS measures capture the correlations of a time-frequency representation of the signal, which turns out to be highly speaker-dependent (Grenier 1977; Gish et al. 1986; Bimbot and Mathan 1993). Here again, the advantages of SOS measures are similar to those of DTW. In particular:

- Training and testing are fully deterministic processes.
- Scores can be interpreted to some extent as logarithms of likelihoods.

Naturally, the correlation matrix of a speaker is better estimated over a relatively long period of speech (several seconds). Nevertheless, reasonable performance can be obtained on segments as short as three seconds. A detailed evaluation of these measures, as well as a further discussion on their possible use as reference methods can be found in Bimbot et al. (1995).

Both DTW and SOS measures are based on totally reproducible algorithmic methods. They probably do not represent the ultimate technological solutions to speaker recognition. However, their systematic use as calibration approaches should at least allow discrimination between relatively trivial tasks,⁵⁴ and those which are really challenging. In that sense, the issue of defining reference systems is a matter of efficiency for research and for evaluation methodology.

11.5.2 Human calibration

In speech technology assessment, human performance in achieving a well defined task is one of the most popular methods to obtain a comparative evaluation for an automatic system. For instance, the classical HENR method to assess speech recognisers (cf. Chapter 10) and the well defined and established set of listening tests (Diagnostic Rhyme Test (DRT), Mean Opinion Score (MOS), etc.) used in assessment of speech output systems (cf. Chapter 12) are well known techniques in speech assessment. Moreover, human calibration may offer an opportunity to investigate the human approach to the solution of the problem, so that the automatic system may take advantage of this knowledge.

In recent years, a significant amount of effort in the field of speaker recognition has been spent on answering the question of how accurate the automatic methods for speaker identification and verification are compared, to the performance of human listeners. From these investigations, the question was raised whether automatic speaker recognition is one area of speech processing where machines can exceed human performance. Unfortunately, as no common formalism has been established and as the experiments reported in the literature usually do not have the same experimental conditions, conclusions are not clear.

In fact, the large number of possible distinct factors that should be managed in a listening session, such as the number of speakers, the duration of voice material, voice familiarity, the phonetic content of the speech material, delay between sessions, etc., make the definition of a standard listening test a very difficult goal. Moreover, the comparison between automatic methods and human listeners can easily end up in the selection of a task reasonable for the automatic system and unfair to listener capability.⁵⁵

Nevertheless, for speaker verification the problem becomes simpler (Rosenberg 1973; Federico 1989), as a human calibration of the task can be done by a series of pair tests, to which the human listener is asked to judge whether they belong to the same speaker or not. A necessary step is now to define and test procedures for listening tests in automatic speaker verification, so that effort in this field will not vanish owing to a lack of reproducibility, or

⁵⁴In particular, those for which any method will work!

⁵⁵Especially for speaker identification when the number of speakers is large.

to multiple test conditions.

Given that listening tests are very time consuming and cost intensive research activities, it is not realistic to envisage such a human calibration on every existing database. A good compromise would be to dedicate some effort to human calibration of standard databases. Nevertheless, further research and experiments in the human test field is necessary in order to fix possible standards and recommendations supported by both theoretical models and experimental results, as the listening methods would surely be helpful both in development and assessment of speaker recognition systems.

11.5.3 Transformation of speech databases

Speaker recognition systems are sensitive to external factors, such as noise, transducer and channel characteristics, etc. Performance is also influenced by intra-speaker variability, i.e. all kinds of *unintentional* or *intentional* modifications of the speaker's voice. Among the most common unintentional modifications are changes in health conditions⁵⁶ and perturbations in speech production caused by the environment, such as surrounding noise.⁵⁷ For what concerns intentional voice modifications, one can distinguish between *voice masking* and *mimicry*, depending on the goal of the speaker.⁵⁸

It is virtually impossible to construct a database which would be representative of all combinations of these factors varying in sufficiently small steps to cover any imaginable situation! The idea behind *indirect assessment* (SAM-A 1993) is twofold: first to develop realistic models of these factors, and then to apply these models to pre-existing databases and vary them in a controlled manner. For each factor, the limits of acceptable variations can be measured, providing a *sensitivity profile* for the system under test.

At our stage of knowledge, it is certainly not possible to develop a viable model for each factor mentioned above. In fact, external factors appear to be easier to simulate than intra-speaker variability. But we believe that indirect assessment through transformations of speech databases is certainly a challenging research topic, the outcomes of which should contribute to simplify assessment procedures considerably.

11.6 Applications, systems and products

The paralinguistic information conveyed by the speech signal is and will be exploited in many practical applications. In some cases, it is used to improve another aspect of speech communication (as in speech recognition and speech synthesis). In other cases (speaker verification/identification, speaker change detection, pathology detection, age/sex/language identification, etc., it is the essence of the application.

⁵⁶ A cold for instance.

 $^{^{57}\,\}mathrm{The}$ so-called Lombard effect.

⁵⁸ Not to be recognised or to be mistaken with somebody else, respectively.

It is unrealistic to review all possible applications, many of which are unknown at this time. What is attempted here is a classification of potential applications with a main focus on speaker verification.

Several approaches to the problem of speaker recognition are now efficient enough to bring advantages to some real-world applications. It is a matter of putting together a reasonably efficient algorithm, a robust decision strategy and a carefully designed ergonomy, in a well-engineered product. As an additional layer of security, voice verification can be effective in several applications, even if not 100 % reliable (personal identification numbers are not 100% secure either). Beside its dissuasive effect, it is clear that it can reduce some significant part of unauthorised access or transactions, without being perceived as a hindrance by the conventional customer. Multi-modality will extend progressively its action field in the everyday world, and voice recognition will undoubtedly have a role to play in this development.

The success of an application depends on many factors including the use of robust algorithms, the realisation of system components to implement these algorithms, the integration of such components in competitive products and the acceptance of these products by users. The points of view of the different actors (users, service providers, system manufacturers and researchers) should be taken into account.

Products are proposed by service providers who rely on system manufacturers to offer generic systems. A limited number of selected systems and products are described to point out some of the factors that could be important to take into account in their design and evaluation.

11.6.1 Terminology

In order to facilitate the comprehension of the following discussion, a number of terms are clarified and an example is given (see also Chapter 2).

- An *application* is a specific problem (for example: a security problem) which could be solved using competitive and/or collaborative technologies.
- A *product* is the integrated hardware and software customised for an application.
- A system is a generic component designed to be integrated in products.

Service providers take the responsibility (and the risk) of installing specific products in response to market demands for applications. They acquire the selected technology from *system manufacturers* and can contract *integrators* to supply the actual products.

Example: A bank (service provider) decides to set up a telephone service to facilitate the transactions of its clients (the application). Such a decision is usually governed by economic factors and compromises cost and client satisfaction. The bank is responsible for the specifications of the product

that will be installed. It may produce a call for proposals describing these specifications. System integrators will answer the call with specific solutions (potential products) made of system components integrated with a user interface. The bank will select the "best" solution.

11.6.2 Typology of applications

It is unrealistic to review all the applications that could benefit from the knowledge (recognition) of speaker characteristics. The overlap with the applications of speech technologies could become more and more important in the future. Techniques for speaker recognition can be used for speaker selection or the adaptation of multi-speaker or speaker-independent speech recognition systems, and can bring improvements to speech processing techniques in general, including synthesis and coding.

- In Automatic Speech Recognition, it may be important to recognise the accent (regional, foreign) of the speaker, his speaking rate, his style, his mood, etc. The recogniser itself could be adapted dynamically to the speaker.
- In Speech Synthesis, the control of voice characteristics offers the possibility of simulating any speaker, of conveying emotion, etc.
- Pronunciation training aids and reeducation software could be adapted to the speaker.
- Speech coders could also adapt to the voice of the talker.

In most of these cases, it may be inappropriate to treat separately linguistic and paralinguistic aspects of the speech signal. The paralinguistic aspects are certainly more important in the following tasks:

- identification or identity verification
- speaker change detection
- speech pathology detection and evaluation

Speaker change detection can be used for automatic speaker labelling in recordings or focussing on the current speaker in video conferences. For some multimedia applications, it may be very useful to access, speaker by speaker, the recording of a conversation or of a radio or television program. Here, the speech to be processed is reasonably sequential, which makes the algorithms usually more efficient, but it may happen that several speakers talk at the same time.

Most of the following discussion will focus on speaker verification. In this context, the most obvious dichotomy separates *remote (telecom) applications* from *local (on-site) applications*. Remote applications are typically performed over the telephone. A major problem with the telephone is the diversity of telephone sets and channel paths. The microphone and the environment is much better controlled with local applications.

11.6.2.1 Telecommunication (remote) applications

One of the most obvious uses of speaker recognition techniques is caller authentication over the telephone network. In the framework of such applications, the main task is therefore speaker verification. The user claims his identity, most of the time by dialling (or saying) a personal code number. Then, either a code word is required to authenticate the speaker, or his utterance of the code number is used for verification purposes. Typical application areas are:

- calling cards with home bill charge
- banking (checking balance, transfer of funds)
- payment by credit card: remote payment by credit card is usually accepted by providing the name of the owner, the card number and the expiration date (all this information is printed on the card)
- teleshopping (grant authorisation to transfer money)
- stock exchange operations (purchasing, selling stocks)
- home incarceration, alcohol rehabilitation program
- military applications
- access to or modification of information on remote servers (restrict access to authorised users).
- access to computing facilities from a remote terminal (coupled with encryption), where passwords are currently used, and alternatives necessitate non-standard equipment (card reader, scanner, electronic pen, encrypted key, ...)

The main applications for speaker verification over the telephone network are of two kinds: the first kind is banking and remote transactions, the second kind is access to licensed databases. It is obvious that both fields do not require the same level of security: it is usually less costly to let someone unauthorised have access to a database than to allow somebody to operate some transaction that can involve large amounts of money. In many countries, it is possible to pay by credit card over the phone, without any other verification of the customer than the consistency between the customer's name, his credit card number and its expiry date (all of them being on the credit card itself!). Some kind of basic speaker verification (for instance user-specific text-dependent verification), even with tolerant thresholds, could certainly bring more security to this type of transaction. Note that, in this case, the speaker characteristics should be centralised in a single place that delivers the transaction authorisations, and would require special equipment for each supplier, but the verification could take place off-line because the confirmation of the transaction does not usually need to be immediate.

11.6.2.2 On-site (local) applications

With on-site applications, the person whose identity is to be verified must be either physically present in one particular location or in direct contact with a device under the control of the service provider. This offers some freedom for the product designer to choose from a variety of techniques. Keys, badges, codes, ... are used most frequently. However, they may not insure a sufficient level of security. In such cases, biometric verifiers offer an alternative (or a complement).

Typical applications are:

- access to or control of equipment,
- access to a secured area (nuclear plant, military premises),
- voice key (home, car),
- mobile telephone, personal assistant (only respond to the voice of his owner),
- Automatic Teller Machine (ATM).

These applications are the equivalent of database access and remote transactions over the telephone. However, the differences to telecommunication applications come from four facts:

- Environment factors and the signal bandwidth can be more easily controlled.
- Automatic verification can send an alarm in case of doubt.
- The customer can carry his voice characteristics with him (on an intelligent card, for instance).
- The voice verification technique can be associated more easily with additional identity verification (multimodal) techniques.

For example, a possible implementation of a voice verification system for money distributors could simply refuse the transaction (or limitate its amount) if the voice characteristics do not match sufficiently the identity corresponding to the Personal Identification Number (PIN). Note that, even if a thief steals a credit card and the PIN is attached to it, he may still not know the voice of the user, and may have difficulty in getting information about it. If he knows the voice of the owner, he may be unable to imitate it. Even if some false acceptances take place, the amount of fraudulent transactions will be necessarily reduced. However, if too many false rejections occur, the bank may lose some of their clients. The probability of an impostor having a voice similar to the user being small, the system can be quite tolerant and still reduce the number of fraudulous withdrawals, be somehow dissuasive to impostors, without offending a significant subset of the regular users. If additional procedures are put into action, such as taking a picture of the user in case of doubt on his identity, or performing some kind of face recognition, dissuasion is reinforced, since it requires a more elaborate and less casual strategy to be resorted to by a possible impostor. For this kind of application, the voice characteristics of the speaker can be stored on the magnetic tape or the chip of his card, which does not require centralised access. The verification must take place in nearly real-time to avoid undesirable delays in the transaction.

In the context of access control, a reasonable system would be based on rather strict verification, with a "call for assistance" procedure in case of rejection.

11.6.3 Examples of speaker verification systems

Technology providers usually offer speaker verification sytems that could be integrated in products. In this section, we examine a few examples.

- VOCALIS Ltd. (at that time part of LOGICA) developed a real-time telephone-based text-dependent system in 1988. The system's decision making logic has four layers of authentication prior to making the final decision on the claimed identity of the speaker. Individual password and PIN are used as the first two layers of authentication. A number of words from pre-defined vocabularies are also chosen randomly at the later stages of verification. The enrolment process follows a robust strategy, in order to guarantee that the reference template is relevant. VOCALIS' speaker verification system is developed for a UNIX platform and makes use of embedded hardware. It is fully integrated into the company's communications infrastructure. An extensive field trial of the system was organised in conjunction with a telephone service company in the US, where over 400 people from four different geographical groups registered with the system. The system allowed users to access US Government databases out of office hours. All registered participants were encouraged to make at least ten calls over a three month period, using as many handsets as possible. The registered speakers were also asked to make two impostor calls and were purposely provided with information to break the first two layers of authentication. At the end of the trial, more than 4000 calls were logged. A detailed analysis of the results showed a 5% equal error rate purely for the verification algorithm. None of the participants managed to break into the system as impostors when all layers of authentication were active. A similar speaker verification system is available also on the Callserver platform from VOCALIS.
- ENSIGMA Ltd. proposes a speaker verification system named VERIFIER (Moody 1991), which runs on the Loughborough AT&T DSP32C telephone board. It works both over telephone line and audio line. ENSIGMA claims full portability of the speaker verification software on other boards. The system uses a new approach to the speaker verification problem. In fact, instead of simply matching a speaker's voice against the user's template, ENSIGMA's system compares the spoken word also against a "general world model". The decision is made taking into account the results of these two comparisons. This method has been evaluated over the tele-

phone line and gives 1% false rejections and 1% false acceptances. The method also seems quite robust to background noise. Enrolment consists of the new user repeating a series of digits; the duration of this phase is about one minute. The access consists of the pronunciation by the user of a list of (random) digits requested by the system. System response takes about 5 seconds. The VERIFIER has many properties that make this system a good candidate as a possible reference hardware system for (text-dependent) speaker verification. The multilingual aspect, or generally speaking the vocabulary problem, may be solved only if procedures to build the "reference world model" used to score utterance similarity is released by the producer.

• The CNAPS of Adaptive Solution Inc. works over the telephone line and gives both speaker verification and speaker identification capabilities (Skinner et al. 1992); averaging the true acceptance and true rejection rates the system reaches 95% of correct responses for speaker identification and 97% for speaker verification. These results refer to a field test where the user pronounces the seven digits of their PIN to access a system; for speaker verification, the user communicates a reference number using DTMF to the receiver before starting to pronounce the seven digits.

Voice Control System offers speaker verification in their speech recognition system over the telephone line. It works on several platforms including PC and VME. By August 1990 (Hunt 1991) the performance was 1% false acceptances and 2% false rejections.

A speaker verification facility is also part of the AT&T HuMaNet teleconferencing system (Berkley and Flanagan 1990) that works via ISDN and uses an AT&T digital signal processor DSP32C board on a Personal Computer.

The main problem which service providers and integrators have to face is to be able to predict what level of performance and user satisfaction they are going to get from a given technological solution.

11.6.4 Examples of speaker verification products

Products are systems integrated into a specific application. Examples are given here to emphasise the benefits for both users and service providers.

Some telephone operators provide a toll free service allowing customers to place a phone call from any terminal. The charge for the call appears later on the phone bill of the user. For such an application, the user must be identified. A personal identification number is usually requested. This information could be entered via the dial pad, but the security of such a protocol is very low (a PIN could be easily stolen).

The SPRINT operator in the USA proposes a successful alternative. It uses speaker verification. Upon a call to the service, the user is requested to say the sequence of digits of his account. Digit recognition and speaker verification is performed on this sequence. In case of doubt, he is then prompted a random sequence of digits for further validation (getting some assurance that he is not playing back some prerecorded speech). He is then allowed to dial the phone number of the person he wants to reach.

The SPRINT operator has registered 1.5 million customers for this service. In such an application, a rather high false acceptance rate is tolerable. Impostors are aware that their voice is being recorded. The introduction of speaker verification proved to be dissuasive against fraud. With a rather small increase in the complexity of the access protocol, a satisfactory level of security was achieved. Of course, a higher degree of security could be reached by letting the user dial by voice. The risk of recognition error (although rather small as recognition could be performed in a speaker dependent mode) would necessitate confirmation and a longer dialogue (paid by the service provider!).

In the UK, the BT home banking trial with the Royal Bank of Scotland used speaker verification technology to provide an additional layer of security to PINs (Personal Identification Number). The concern of offending as few genuine customers as possible was of prime importance to the bank.

These examples show that the benefit of a speaker verification system is obtained as a compromise between the reduction of fraud and user acceptability. In particular, it is clear that the equal error rate usually does not correspond to a realistic functioning condition.

11.6.5 Alternative techniques

When choosing adequate technologies to be integrated in a product all alternatives must be taken into account. Security may be achieved in several ways, and voice is only one of the possibilities. We give here a brief overview of some alternative techniques.

The physical support for protecting information or access to unauthorised persons are usually classified into three main families: physical-object elements (key, badge, etc.), information elements (password, combination, etc.) and physical-personal elements (fingerprint, voice, DNA, etc.). The ensemble of systems based on this last family are called "biometric verifier" as they are directly tied to some biological characteristics of the (authorised) person.

The biometric family is divided into two subfamilies: one based on physical characteristics of the person that do not change (unless they are crude events) in the range of few years; the other based on behaviour character-

istics that may change in relation to humour, environment, physical state, and so forth. Because of the high rate of variability of the latter subfamily, corresponding systems are quite difficult to obtain, especially in automatic mode, without expert supervision.

Just to clarify these two subfamilies of biometric verifier let us consider the list of biometric verifier shown in Table 11.7:

Table 11.7: Biometric verifier

Physical	Behaviour
fingerprint	voice
retinal scan	$\operatorname{signature}$
face conform	writing
genetic analysis	cardiac rhythm
hand structure	

There is no unique strategy to select the most appropriate biometric verifier for a given task. However, there are a number of properties that must always be considered in designing a security system. Table 11.8 gives some important ones.

Table 11.8: Important properties for designing a security system

error(s) rates	execution swiftness	access facility
network capability	number of potential users	rate of test per time unit
acceptability	integration facility	$\cos t$

Generally speaking, if we are interested only in performance, a broad table summarising the value for different biometric verifiers may be the one reported by Peckham (1990) (see Table 11.9).

Modality	False Rejection	False Acceptance
voice (high quality)	1.0~%	0.1~%
voice (telephone quality)	3.3~%	0.4~%
retinal scan	2.8~%	0.0~%
signature	0.2~%	0.6~%
veincheck	$< 1.0 \ \%$	$< 1.0 \ \%$
hand shape	0.1~%	0.1 %

Table 11.9: Values for different biometric verifiers

The same report gives the example of field tests carried out for the US government on speech, signature and fingerprint. These resulted in a recommendation to use speech. Note also that the alternative techniques listed above are more appropriate for on-site applications and that speech remains a unique biometrical identification feature over telecommunication channels, at least until videophone and ISDN become widespread. Even so, some of the alternative techniques could only be implemented through individual sensors, an essential limitation of their applicability.

11.6.6 Conclusions

From the examples above, it is quite evident that voice verification can represent an advantageous additional layer for user authentication. It is also clear that the success of such an implementation relies on an adequate design and engineering of the application, including multiple layer decisions, error recovery strategies and ergonomical considerations. At this stage, additional alternative techniques can also be integrated, for increased security. The future outcome of voice verification techniques from a commercial point of view is certainly strongly connected to requirements in application design.

11.7 System and product assessment

The need for system and product assessment has multiple aspects. The most obvious one is the need that service providers communicate to system manufacturers, as they want to have a clear idea of the benefit that they can expect from some particular technology in the context of the product that they intend to develop and supply. A first question is therefore to define and measure appropriate quantities from which a service provider will be able to predict with reasonable accuracy the consequences of the new technology for its service.

Some form of assessment is also necessary to inform and protect end users. Let us illustrate our argument with the extreme example of a hypothetical "lie detector", a system which would supposedly be capable of detecting when a speaker is purposely not telling the truth. Should such systems be implemented and sold as commercial products without any objective evaluation of these products having been conducted?

For some products, a more or less independent standardisation institute has to deliver formal approval prior to commercial exploitation. For others, the product has to fulfil a certain number of specifications before it can be sold. At the moment there is no such procedure concerning speaker recognition technology.

It is part of the responsibility of the scientific community to provide tools and methods in order to assess the performance of systems.

The rest of this section lists the multiple aspects that have to be taken into

account. However, no commonly agreed standard exists today concerning the relevant quantities to measure and *a fortiori* the methods for measuring them.

Guidelines and recommendations should also be made about the feasibility of applications: it may happen that the general public is not able to tell easily whether a given product is realistic or whether it remains science fiction. While scientists should be aware of new needs, they should also keep the general public permanently informed of the progress and the limitations of our current knowledge.

11.7.1 System assessment

Several commercial systems using speaker recognition techniques are now available on the market, mostly for speaker verification purposes. As with existing algorithms, it is quite difficult to compare them objectively. This fact stresses the need of standard evaluation procedures these systems: one possible approach is the definition of standard applications. Such a procedure should specify very accurately the nature of the applications, and integrate the speaker verification layer into a predefined structure of authentication layers. It is indeed clear from several examples, that the error rate on the pure voice verification algorithm can have very different overall consequences over the global score of an application, depending on the way voice verification is integrated with other verification layers. A drawback of such an approach is that the applications to which a system is targeted may be quite different from reference applications, and the evaluation figures may become somehow meaningless.

Performance is only one aspect of system assessment, and service providers will consider a number of other aspects:

- initial and maintenance costs,
- scalability and upgradability,
- easiness to incorporate improvements or modifications,
- availability of text-dependent, text-prompted and text-independent algorithms,
- flexibility of user enrolment,
- adaptability of speaker models over time ("batch adaptation"),
- control over the decision strategy,
- estimated durability and robustness,
- trustworthiness of the manufacturer,
- ergonomy of the man-machine interface,

Most of the aspects mentioned in this list are not possible to quantify. They are elements in the judgment that the service provider will take into account when deciding to retain a given technological solution for his application.

11.7.2 Product assessment

The success of a product could, to some extent, be measured *a posteriori*. Some of the relevant figures are:

- number of units installed,
- benefit to the service provider,
- number of registered users,
- evolution of user registration over time,
- number of successful access from users,
- average time for a successful access,
- evolution of the access rate over time,
- actual false acceptance and false rejection rates (if available),
- estimation of fraud reduction,
- number of complaints from users,

It may prove difficult to obtain some of these figures. In the context of secure access (either local or remote), one important aspect is the actual reduction of fraud. There is a cost/benefit tradeoff to examine.

The installation of a speaker verification system can be expected to have an impact on the population profile of the users. In particular, it can:

- dissuade impostors (afraid to get their voice recorded),
- discourage clients (disappointed by extra difficulties in accessing a service),
- encourage new clients to register for the service.

A service provider would indeed be much more interested in tools and guidelines to permit an *a priori* assessment. He must analyse the user needs and/or the cost/benefit tradeoffs of the application from (sometimes) conflicting perspectives. He could obtain some indications from an analysis of existing applications. He should compare alternatives initially on paper and eventually with simulations. In some situations (telephone services are good examples), a "Wizard of Oz" (WOZ) technique (see Chapters 3, 9, 13) could be set up to investigate the man-machine interface and the user acceptability. With such a technique, a real or simulated application is set up, but all decisions concerning speech recognition and/or user verification are taken by a human operator ("the man behind the curtain"). Fuzziness on these decisions could be introduced to simulate recognition errors.

Once the initial specifications of the product are set up, a prototype is realised and field trials are organised. This is a necessary step before launching a large scale application. The specifications should be flexible enough for possible readjustments. The prototype should allow for the monitoring (or at least the recording) of the user interaction. These recordings are used to fine tune the system.

The length of the speech material necessary for training and testing, the false rejection rate and the ergonomy of the implementation will influence user
acceptability. For example, experimental evaluations indicate that some speakers ("goats") may be rejected quite often. The test population should be large enough to show the distribution of false rejection. In fact, the overall false rejection rate of a system may not help determining its user acceptability. For instance, a system that rejects systematically 5% of the user population but accepts with no errors the remaining 95% may be better accepted than a system with a 5% false rejection distributed uniformly on the entire population. Therefore it may be more meaningful to evaluate the proportion of the client population for which the false acceptance rate is under a certain threshold, rather than the average false acceptance rate as is commonly done. To satisfy the "goat" population, a backup strategy (using a human operator for example) can be proposed.

With some applications (like banking), imposture could be discovered *a posteriori*. The system could record all the transactions, and the voice recordings could be used to trace the impostor. Legal issues related to these recordings and their use should be investigated carefully.

To conclude, voice verification technology is certainly at a stage where its future use is partly conditioned to the ability of the system manufacturers to convince the service providers and the end users of the advantages that can be gained from such techniques. This conviction can only be reinforced with the definition of well-established assessment tools.

11.7.3 Recommendations

The intent here is to provide guidelines for actors in the technical and economic chain, from algorithms to products, who are concerned with speaker recognition. These actors are research scientists, development engineers, equipment manufacturers, service providers and end users.

- 1. Research scientists should investigate the basis of biometric identity verification. This fundamental research will eventually suggest new approaches. They are also responsible for the specification of methods and databases to evaluate these approaches. Part of the research should be rather applied to demonstrate that fundamental results will benefit society. Scientists should be well aware of new needs of our society and should suggest solutions.
- 2. Development engineers are in charge of the integration of algorithms on real-time platforms. They should be well acquainted with results obtained under laboratory conditions. They should interact with researchers to scale the implementation aspects of algorithms. They usually work for equipment manufacturers.
- 3. Equipment manufacturers need to select the best algorithms and platforms to offer competitive systems. They must be well aware of market trends. They should provide evidence and results of case studies to help predictive assessment. They must claim results on public databases. They should offer demonstrations and make systems available to facilitate evaluation.

- 4. Service providers are responsible for products. They should compare alternative technologies to solve problems. They must carefully design the ergonomy of the man-machine interface and run field tests.
- 5. End users will finally decide the success of applications.

11.8 Forensic applications

One important field in which speaker recognition is usually involved is the field of forensic applications. This concerns cases for which victims or witnesses have an auditory memory of the criminal voice and cases for which recordings are available. In the case of recordings, the content of the tapes may be important (Who is speaking? What is being said? Which sounds could be identified? ...). The tapes should usually be analysed to check whether or not they have been manipulated, erased, recorded over, etc. The signal may require enhancement to be comprehensible. In some cases, the voice of a suspect should be compared with a recorded voice. This is a very special case of speaker verification with many specific difficulties:

- The recordings are non-contemporaneous.
- The original recording could be noisy, distorted, manipulated, etc.
- The original recording could be of short duration.
- The voices could be disguised.
- The texts are usually different.
- The speakers are not cooperative.

Various methodologies for approaching this problem have been proposed; they may substantially be classified into the following three categories according to the criterion adopted for the analysis and the characterisation of the voice signal: the listener method, the spectrographic method and the semi-automatic method.

11.8.1 Listener method

In the listener method, the decision of similarity is taken by human experts after audition of speech samples. One of the methods is the repeated listening of the available samples by a group of experts looking for similarities in linguistic, phonetic and acoustic features. The voices listening test can be done either sequentially or alternatively or finally, by simultaneous presentation of the voices, if allowed by the instrumentation. The acoustic material is carefully prepared, selecting from the database suitable utterances of the unknown and of the reference speaker, eventually augmented with voice samples coming from speakers with the same acoustic-phonetic characteristics. These utterances are then compared two by two, inverting pairs of voices from the same speaker. Studies have been carried out to investigate how well listeners can recognise speakers (Stevens et al. 1968). Listener performance is a function of acoustic variables such as the signal-to-noise ratio,

the speech bandwidth, the amount of speech material, complex distortions of the speech signal introduced by speech coding transmission systems, etc. Several experiments using different texts for comparing and evaluating the performance under different degraded conditions showed that human listeners are robust speaker recognisers when presented with degraded speech. This is due to the fact that there are many sources of knowledge that contribute in various ways to speaker recognition, providing weak, moderate or high discrimination power. However, as with any human decision process, it must be stressed that the listener method leads to a subjective decision. Nevertheless, the listener method is still used in some countries for forensic applications as a technique for speaker recognition.

11.8.2 Spectrographic method

The spectrographic method for speaker recognition makes use of an instrument that converts the speech signal into a visual display. For many years the reference instrument was the "Voice Identification Inc., Sound Spectrograph, model 700". This instrument is able to give a permanent record of changing energy-frequency distribution of a speech wave through time. Usually, the frequency range is 0–4000Hz, and the bandfilter is 300Hz. Since spectrograms are visual representations of the speech signal, they convey information about the message spoken by the speaker as well as about the speaker himself. For this reason, these patterns were thought to be used as a way of identifying speakers. For example, when the recordings of the voice of two individuals are obtained, an examiner may be able to give an opinion about the similarity between two recordings, if there are common phonetic elements between their speech recordings. This method for speaker identification was originally proposed in Gray and Kopp (1944), but its use for forensic applications was not considered until 1962, when Kersta (1962) published the results of experiments on one-word spectral comparison in closed-set tests. Further studies were also carried out by Stevens et al. (1968) and by Tosi et al. (1972), who presented the results of research at Michigan State University on the basis of a "forensic model" with openset tests. These results have been analysed from Bolt (1970), by observing the error rates of false identification and of false elimination. They observed that the error rate is dependent upon a lot of factors, i.e. different conditions of environment noise, the change in the psychological state of the speaker, his attempts to alter his voice, the recording conditions, the orthophonic or telephonic voice of the talker etc.; in particular, the error rate is widely dependent on the examiner and is increased by changing from trained to untrained examiners. Owing to these factors and to other restrictive conditions that affect the error rate of examiners, this method is today of no great interest to scientists in speaker recognition tasks.

Both the listening and the spectrographic methods are subjective techniques

based, the first one on aural comparison of recordings and the second one on visual examination of spectrograms in order to attribute two voice samples to the same talker. These decisions are taken by one or several experts, according to some process that is clearly impossible to formulate. Moreover, the process that leads to attributing to a particular individual the quality of "expert" is far from being calibrated itself. Very often also, experts are asked to give a probability of confidence on their own judgments. Are forensic speech experts submitted to a benchmark test before they are recruited? Does it make sense to give a figure measuring one's own selfconfidence on one's own decision? There are certainly several issues about forensic applications of human speaker recognition that have to be called into question.

11.8.3 Semi-automatic method

It is clear that some of the research carried out in automatic speaker recognition is targeted towards forensic applications. Most of the methods described in the previous chapter could be adapted to this goal, especially text-independent ones. However, if they offer a certain rate of success (under restrictive conditions), it must be stated clearly that they are far from being faultless, and that the consequence of an erroneous decision can be dramatic. Actually, since some expertises use very basic statistical methods (such as mean and variance computation), there does not seem to be any elaborate speaker recognition system publicly used for forensic applications, but the possibility is not excluded given the degree of secrecy surrounding such applications. Nor is it excluded that experts resort to such systems for making their own decisions, with or without saying it explicitly. It is clear that such systems, whatever level they are used at, should be submitted to objective test protocols designed in agreement with members of the scientific community. Frequent contacts between speech researchers and members of the judiciary and the police would also certainly help to clarify the possibilities and the limits to the use of speech and of speech technology in forensic applications.

11.8.4 Recommendations

For many applications of speaker recognition techniques, a wrong decision has only material consequences. For forensic applications, much more serious aspects are involved.

No serious scientific protocol has been able so far to evidence the existence of a fixed, robust, non-modifiable, individual voice characteristic that could be extracted from a speech signal and indicate without doubt the speaker's identity. Therefore, we recommend substituting the term *voice signature* for the frequently used *voice print*: The former renders better the idea of variability and intentionality, while the latter is misleading as it suggests an

analogy with finger prints.

Speech databases, specific to the forensic area, should be carefully designed, recorded and distributed. Typical forensic situations should be simulated, including cases of voice disguise. These databases would consist for instance of a large collection of pairs of voice samples, sometimes from the same speaker, sometimes from different speakers. One of the recordings could take place in a studio, while the other would be submitted to several kinds of environmental, channel, artificial and intentional distortions. It would be calibrated in such a way that the average auditor cannot do better than a random choice. Such databases would be extremely precious for several purposes. One would be to evaluate forensic speech experts and clarify if they can really do much better than an average auditor. It would also clearly evidence to the forensic professionals what the limitations of human and automatic techniques of speaker recognition for such applications are.

11.9 Conclusions

Among all the considerations we have presented here, some may have higher priority than others, but the definition of such priorities cannot be decided on without intensive discussion with many other specialists in the field. We believe however that all of these topics should be taken into account, because they represent different aspects of a same problem, and they should not be addressed independently. We hope that we have laid significant foundations for further discussion and investigation in the field of speaker recognition assessment methodology, and we strongly recommend that the need for such investigations be clearly acknowledged by concrete support from institutions concerned, so that it can be carried on.

12 Assessment of synthesis systems

12.1 Introduction

12.1.1 What are speech output systems?

By a speech output system we mean some artifact, whether a dedicated machine or a computer programme that produces signals that are intended to be functionally equivalent to speech produced by humans. At the present time speech output systems generally produce audio signals only, but laboratory systems are being developed that supplement the audio signal with the visual image of the (artificial) talker's face (cf. Benoît 1991; Benoît et al. 1992).¹ Audio-visual (or: bi-modal) speech output is more intelligible than audio-only output, especially when the audio channel is of degraded quality. In this chapter we will not be concerned with bi- or multimodal speech output systems, and concentrate on audio-only output instead.

We exclude from the domain of speech output systems such devices as tape recorders and other, more advanced, systems that output speech on the basis of complete, pre-stored messages ("canned speech" or "copy synthesis"), irrespective of the type of coding or information compression used to save storage space. We crucially limit our definition to systems that allow the generation of novel messages, either from scratch (i.e. entirely by rule) or by recombining shorter pre-stored units. This definition also includes hybrid synthesis systems where individually stored words (e.g. digits) are substituted in information slots in a carrier sentence (e.g. in time-table consultation services).

It seems to us that two basic types of speech output systems have to be distinguished on the basis of their input, namely text-to-speech (TTS) and concept-to-speech (CTS). Other, more complex, systems combine characteristics of these two.

• TEXT-TO-SPEECH. The majority of speech output systems is driven by text input. These systems convert text printed in normal orthography (generally stored in a computer memory as ASCII codes) to speech. Conventional spelling provides a reasonable indication of what sounds and words have to be output, but typically underrepresents prosodic properties of the message, such as the positions of accents, speech melody, and temporal organisation, including speech rhythm. The prosody of an utterance reflects, among other things, the communicative intentions of the writer of the input text, which cannot be reconstructed from the text alone

 $^{^{1}}$ Moreover, systems have been developed that generate audio output together with visual output (text printed on computer screen, for foreign language learning purposes) and/or tactile output (braille, specifically for the visually handicapped). In such cases the term multimodal output is often used.

- note the title of a much-cited article: "Accent is predictable, if you're a mind reader" (Bolinger 1972). The reconstruction of the writer's intentions is an implicit part of the so-called linguistic interface, i.e. the first part of most advanced text-to-speech systems. All errors in the linguistic interface may detract from the quality of the output speech, and are therefore a legitimate object of evaluation.

- CONCEPT-TO-SPEECH. In other types of speech output systems, especially dialogue systems, the communicative intentions are fully specified at the input stage: the system itself determines what message it wants to get across. It may still be the case, of course, that the dialogue system has misconstrued a user's request, and consequently issues an inappropriate response message, but this should not be considered an error on the part of the output system.
- INTERPRETING (OR TRANSLATING) TELEPHONY (SL-TRANS, cf. Morimoto et al. (1990); JANUS, cf. Waibel et al. (1991)) and face-to-face spoken dialogue translation (VERBMOBIL, cf. Wahlster 1993) combine characteristics of both TTS and CTS. Interpreting telephony, for instance, a spoken utterance in one language (e.g. Japanese) is decomposed into its linguistic message and its speaker-specific properties (e.g. voice characteristics, speed, pitch range). The linguistic message is converted to text, and transmitted. At the receiver end the text is automatically translated into another language (e.g. German) and then converted back to speech in the target language setting the synthesiser's speaker specific parameters such that the personal characteristics of the source speaker are approximated in the output signal. Crucially, the sender's intentions do not have to be inferred from the textual representation of the message; the intended focus distribution can be reconstructed directly from the properties of the source language speech signal.

12.1.2 Why speech output assessment?

In spite of the rapid progress that is being made in the field of speech technology, any speech output system available today can still be spotted for what it is: non-human, a machine. Most older systems will fail immediately due to their robot-like melody and garbled vowels and consonants. Other, more recently developed synthesis methods using short-segment waveform concatenation techniques such as PSOLA (Moulines and Charpentier 1990) yield segmental quality that is very close to human speech (Portele et al. 1994), but still suffer from noticeable defects in matters of melody and timing.

As long as synthetic speech is inferior to human speech, speech output assessment will be a major concern. Speech technology development today is typically evaluation-driven. Large scale speech technology programmes have been launched both in the United States and in Europe (for overviews see O'Malley and Caisse 1987; Van Bezooijen and Pols 1989; Pols 1991). Especially in the European Union, with its many official languages, a strong need was felt for output quality assessment methods and standards that can be applied across languages. With this goal in mind the multinational EU-ESPRIT SAM project was set up (Fourcin et al. 1989), and later the EU Expert Advisory Group on Language Engineering Standards (EAGLES) programme started; both initiatives included a working group on speech output assessment.

Speech output assessment may be of crucial importance to two interested parties, the systems designers and developers on the one hand, and the prospective buyers and end users of the system (possibly represented by consumer organisations) on the other.

- Developers are intent on improving their speech output systems. However, designers who have grown up with their system are used to all its habits; they are likely to understand its output much better than first-time users, and will overrate its performance level. Less subjective quality assessment techniques are needed in order to determine how well a system performs relative to a benchmark test, or how favourably it compares with a previous edition of the system or with other designers' systems (*comparative testing* or *performance evaluation*). To the extent that a system performs less than perfect, the designer will have to learn which aspect(s) and/or components of the system are flawed. Designers will therefore also be interested in *diagnostic evaluation*, either by doing detailed error analyses on the test results, or by running component-specific tests.
- The needs of *systems users* (end users and/or systems providers) are different than those of designers but they, too, rely heavily on assessment techniques. Prospective buyers will always have a specific use of their speech output system in mind. Understandably, they will want the simplest, and therefore cheapest, system that satisfies their needs. The buyer (or his consumer organisation) will therefore need an absolute yardstick in order to determine beforehand if the output speech is good enough to get the message across in the given application.

12.1.3 Users of this chapter

This chapter serves the potential needs of several disparate groups of readers. Since not all parts of this chapter will be equally relevant to every reader group, we will identify reader groups, and point out which parts of this chapter, after sections 1–3, have particular relevance for each group. The reader groups, of course, overlap to a large degree with the parties interested in speech output evaluation discussed above, but a more refined classification seems to be in order. We will distinguish the following groups of readers:

• Developers of speech output tests

Speech output assessment is an expanding field. New tests become available in rapid succession, so that test developers want to keep abreast with

what is new. Test developers will want to know the advantages and disadvantages of tests proposed in the literature, and need to know what requirements will be made to the next generation of tests. This chapter discusses many alternative tests and testing methodologies, and makes recommendations as to what type of tests are more suited for a specific purpose. The chapter will also indicate which direction speech output testing should take in order to meet the testing requirements of the next generation of speech output systems (Section 12.6).

Readers in this group will be interested in Section 12.5 on glass box (diagnostic) testing as well as in Section 12.4 on black box testing. When systems are in their early developmental stages, glass box testing will be most relevant; when systems have sufficiently matured, black box tests are in order. Field testing (Section 12.4.2) will generally be deferred until the end user groups and their specific applications are known. Field testing will often be conducted by, or at least in close cooperation with, systems procurers for end users (see below).

The chapter will be especially useful to those readers who are not test developers at this time but aspire at becoming test developers in the near future. These readers include, of course, students at the Ph.D.-level who want to make a career in speech output testing. A second group of readers who are new to the field are developers of speech technology in Eastern Europe and in certain third world countries where computer technology is now widely available at affordable prices, generating an immediate need for the development of speech output systems and tests in the languages of the areas concerned.

Developers of speech output systems

A lot of speech technology research and development takes place in small high-tech companies. The research staffs are often too small to warrant the appointment of a fulltime test evaluation expert, so that a lot of diagnostic do-it-yourself testing is going on. Developers who are not evaluation experts themselves will find this chapter a useful source of information. It identifies standard tests and test suites that are readily available for a range of diagnostic purposes (see the end of this chapter and the appendices). The remarks made above with respect to newcomers to the evaluation field apply here as well. This type of reader should concentrate on the glass box approach (Section 12.5).

PROCURERS OF SPEECH OUTPUT SYSTEMS

At the most user-oriented end of the spectrum, procurers of systems will find our chapter of interest. Procurers, who themselves will more often than not be inexperienced to the field of speech output technology, will not normally be interested in diagnostic testing. They will be looking for a single figure of merit on the basis of which to decide on the system that is best for a given range of applications. This reader group is most difficult to deal with since their needs are most divergent. There are no offthe-shelf tests that satisfy their needs. Rather we will provide numerous examples which may serve as guidelines on how to go about field testing speech output systems for specific applications. This type of reader should concentrate on those parts of this chapter dealing with black box output testing and field tests (Section 12.4).

Given this division of reader groups we will present two types of recommendation, if and when we can.

- The first type suggests what decisions can be made in the present situation with what is available today, or can be made available with little effort in the immediate future. These recommendations will be found throughout this chapter, in concise format and numbered.
- The second type of recommendation that we will be making, outlines possible courses of test development for the mid and long term. Such recommendations, predominantly aimed at the evaluation experts, will be presented in Section 12.6 at the end of this chapter, in less explicit format; they will not be numbered.

12.2 Towards a taxonomy of assessment tasks and techniques

To structure our overview of TTS assessment tests we will discuss a number of useful distinguishing parameters, which partly overlap with earlier attempted taxonomies (see e.g. Van Bezooijen and Pols 1990; Pols 1991; Jekosch and Pols 1994; Goldstein 1995) and explain the relationships among them, before dealing with any specific assessment techniques. Figure 12.1 illustrates the relationships among the various dichotomies that make up our taxonomy. It will be apparent from the figure that the dichotomies are non-orthogonal. For instance, tests that have been developed to evaluate specific modules within a speech output system (glass box approach), will only be used in a laboratory setting. The subdivision of human listener tasks will only be relevant when tests involve human listeners; therefore no task variables will be specified under automated test techniques. In the subsequent paragraphs we will outline and discuss the taxonomy of speech output assessment techniques following as much as possible the structure of Figure 12.1. Note that the dichotomies used are intended as contrastive dimensions so as to create a multidimensional space within which speech output tests can be located. The terms involved in a dichotomy should not be construed as labels identifying mutually exclusive approaches to speech output evaluation.

12.2.1 Glass box vs. black box

Text-to-speech systems generally comprise a range of modules that take care of specific tasks. The first module (or complex of modules) converts an orthographic input string to some abstract linguistic code that is explicit in its representation of sounds and prosodic markers. Various modules then act upon this symbolic representation. Typically, one module concatenates



Figure 12.1: Relationships among dimensions involved in a taxonomy of speech output evaluation methods

Any path from the root down to any terminal that does not cross a horizontal gap, constitutes a meaningful combination of test attributes.

the primitive building blocks (phonemes, diphones) in their appropriate order, another implements what coarticulation is needed to obtain smooth human-like transitions between successive building blocks. Prosodic modules, taking the positions of word stresses, sentence accents, phrasal and sentence boundaries into account, are then called upon in order to provide an appropriate temporal organisation (local accelerations and decelerations, pauses) and speech melody.

End users will typically be interested in the performance of a system as a whole. They will consider the system as a *black box* that accepts text and outputs speech, a monolith without any internal structure. For them it is only the quality of the output speech that matters. In this way systems developed by different manufacturers can be compared or the improvement of one system relative to an earlier version can be traced over time (*comparative testing*). However, if the output is less than optimal it will not be possible to pinpoint the exact module or modules that caused the problem. For *diagnostic* purposes, therefore, designers often set up (*glass box.* evaluations with experimental character. Keeping the effects of all modules but one constant, and systematically varying the characteristics of the free module, any difference in the assessment of the system's output can be attributed to the variations in the target module. Glass box testing, of course, presupposes that the researcher has control over the input and output of each individual module.

12.2.1.1 Recommendations on choice of test methodology

- 1. Use a glass box approach if you want diagnostics in order to improve your speech output system.
- 2. Use a black box approach if you want to assess the overall performance of speech output systems.

The dichotomy between glass box and black box testing is basic to speech output testing, which has led some researchers to propose a strict terminological division whereby "evaluation" signifies glass box testing (or: *diagnostic evaluation*) only, and "assessment" is reserved exclusively for black box testing (or: *performance evaluation*). In this chapter we will use the terms, "testing", "evaluation" and "assessment" indiscriminately, and use disambiguating adjectives whenever there is a risk of confusion.

12.2.2 Laboratory vs. field

Ideally, any speech output system should perform at the same level of adequacy as a human speaker (though there might be counterexamples where users want to know whether they are being addressed by a machine or a human; in such applications is seems advisable to use synthetic speech that is recognisably non-human). Such a system would be optimal for any application. However, given that systems available today are less than optimal, it is important to know which aspects of a system's performance are essential to a specific application. Speech output systems typically form an element of a larger human-machine interface in an application with a specific, dedicated task. In practice this means that, quite probably, the vocabulary and types of information exchanges are restricted and domain-specific, so that situational redundancy is likely to make up for poor intelligibility. On the other hand, speech output systems will often be used in complex information processing tasks, so that the listener has only limited resources available for attending to the speech input. Also, end users may have different attitudes towards, and motivations for, working with artificial speech than subjects in laboratory experiments, especially when the latter have not been explicitly selected so as to be fully representative of the end users. It is often hazardous, therefore, to predict beforehand, on the basis of laboratory tests, how successful a speech output system will be in the practical application. Generally, as an application situation contains more specific aspects, less prediction of field performance is afforded by laboratory tests. Output systems will have to be tested in the field, i.e. in the real situation, with the

real users. The use of *field tests* will be limited to one system in one specific application; results of a field test cannot, as a rule, be generalised to other systems and/or other applications.

12.2.2.1 Recommendations on choice of test methodology

- 1. Do not rely on laboratory tests alone.
- 2. As soon as there is a discrepancy between the laboratory setting and the true field situation (in terms of environment, tasks, type of listener) field testing is necessary.

12.2.3 Linguistic vs. acoustic

The more complex TTS systems can roughly be divided into a linguistic interface that transforms spelling into an abstract phonological code, and an acoustical interface that transduces this symbolic representation to an audible waveform. The quality of the intermediary representation can be tested directly at the symbolic-linguistic level or indirectly at the level of the *acoustic* output. Testing the audio output has the advantage that only errors in the symbolic representation that have audible consequences, will affect the evaluation. The disadvantage of audio testing is that it involves the use of human listeners, and is therefore costly and time-consuming. Moreover, the results of acoustic testing are unspecific in that the designer is not informed whether the problems originate at the linguistic or at the acoustic level. As an alternative the intermediate representations in the linguistic interface are often evaluated at the symbolic level. It is, of course, a relatively easy task to compare the symbolic output of a linguistic module with some pre-stored key or model representation and determine the discrepancies, and this is what is normally done. The non-trivial problem is where to obtain the model representations. These will generally have to be compiled manually (or semi-automatically at best), and often involve multiple correct solutions.

12.2.4 Human vs. automated

In the majority of test procedures, *human subjects* are called upon in order to determine the quality of a speech output system. This should come as no surprise to us, since the end user of a speech output system is a human listener. However, there are certain drawbacks inherent in the use of human subjects. Firstly, humans, whether acting as single individuals or collectively as a group, are always somewhat noisy, i.e. inconsistent, in their judgments or task performance; the results of tests involving human responses are never perfectly reliable in the statistical, psychometric sense of the word. Another drawback of tests involving human subjects it that they are time-consuming and therefore expensive to run. Recent developments, which are still very much in the laboratory stage, seek to replace human evaluation by automated assessment of speech output systems or modules thereof. Attempts can be (and in fact have been) made to automatically measure the discrepancy in acoustical terms between a system's output and the speech of the human speaker that serves as the model the system is intended to imitate. This is the type of evaluation technique that one would ultimately want to come up with: the use of human listeners is avoided, so that perfectly reproducible noiseless results can be obtained in as little time as it takes a computer to execute the program. At the same time, however, it will be clear that implementation of such techniques as a substitute for human listeners presupposes that we know exactly how human listeners evaluate differences between two realisations of the same linguistic message. Unfortunately, this type of knowledge is largely lacking at the moment; filling the gap would be a research priority. Nevertheless, preliminary automatic comparisons of synthetic and human speech output have been undertaken in the fields of melody and pause distribution (Barry et al. 1989), long term average spectral characteristics (Pavlovic et al. 1991) and dynamics of speech in the frequency and time domains (Houtgast and Verhave 1991, 1992). Generally, the results obtained through these techniques show sufficient promise to warrant extension of their scope.² We will come back to the possibilities of automated testing in Section 12.6.

12.2.5 Judgment vs. functional testing

By *judgment testing* (also called *opinion testing* in telecommunication research) we mean a procedure whereby a group of listeners is asked to judge the performance of a speech output system along a number of rating scales. The scales are typically bi-polar adjectives that allow the listeners to express the quality of the output system along a more global or more specific aspect of its performance. Although the construction of an appropriate scaling instrument is by no means a trivial task, a scaling test can be administered with little effort and yields a lot of potentially useful information.

At the other extreme the speech output can be assessed in terms of how well it actually performs its communicative purpose. This is called *functional testing*. For instance, if we want to know to what extent the output speech is intelligible, we may prefer to measure its intelligibility not by asking listeners how intelligible they think the speech is, but by determining, for instance, whether listeners correctly identify the sounds. Consider, as an example on

 $^{^{2}}$ This intended use of objective quality measurement bears a superficial resemblance to the work by Kryter (1962a,b) and associates on the development of the so-called Articulation Index (AI). The AI accurately predicts the loss in intelligibility of speech due to characteristics of the transmission channel by automatically comparing the longterm average spectral characteristics of the input and output speech. However, no AIbased approach can be used to discriminate between the desirable (speech-like) and the undesirable (noise, distortion) features of the input speech.

a higher level of communication, the assessment of an information system using speech output. We may ask users to judge the output quality, but we may also functionally determine the system's adequacy by looking at task completion: how often and how efficiently do the users get the information from the system that they need?

One would hope that the results of judgment and functional assessments converge. Obviously, one would like to use the results of functional assessments in order to gauge the validity of judgments, rather than the other way about. As far as we have been able to ascertain, there has been little research into this matter. Yet, there is at least one set of intersubjective and functional data that was collected for the same group of listeners and stimuli, testing two different text-to-speech systems at three different points in time, from which it appeared that the scaling results were highly correlated with the corresponding functional test scores (Pavlovic et al. 1990).

12.2.6 Global vs. analytic assessment

In a sense there is only one ultimate criterion that determines the quality of a speech output system, viz. its overall quality within a given application. Judgment tests usually include one or more rating scales covering such *global* aspects as "overall quality", "naturalness" and "acceptability". A functional approach to global assessment would be to determine whether users of speech output, when given the choice, choose to work with a machine or with the human original the machine is intended to simulate. Or one may determine if the information exchange is as successful in machineto-human as it is in human-to-human situations.

On the other hand, one may be interested in determining the quality of specific aspects of a speech output system, in an *analytic* listening mode, where listeners are requested to pay particular attention to selected aspects of the speech output. Again, both judgment and functional tests can and have been designed addressing the quality of specific aspects of a speech output system. Listeners may be asked, for instance, to rate the clarity of vowels and consonants, the appropriateness of stresses and accents, pleasantness of voice quality, and tempo. Functional tests have been designed to test the intelligibility of individual sounds (phoneme monitoring), of combinations of sounds (syllable monitoring), of whole words (word monitoring) in isolation as well as in various types of context (e.g. Nusbaum et al. 1986; Ralston et al. 1991).

12.3 Methodology

The great majority of speech output assessment techniques use listening experiments involving human subjects, i.e. functional and/or judgment tests

of speech output at the acoustic level. In the following subsections we will discuss a number of methodological issues that are relevant especially to this type of testing. The issues concern the choice of subjects, test procedures, benchmarks and reference conditions, and precautions to ensure cross-language comparability. Although there is no *a priori* reason why this should be so, no accepted methodology seems to exist for other types of speech output evaluation techniques. As will be obvious from later sections, for example, no accepted methodology can be identified in the field of output evaluation at the symbolic linguistic level. It is unclear in this area what kinds of textual materials should be used in tests, what error categories should be distinguished, and what scoring procedures should be used. We will therefore limit the methodological discussion to acoustic output testing techniques involving human listeners.

12.3.1 Subjects

One of the most important aspects of a measuring instrument is its reliability. How reliable, for example, is subjects' performance in functional intelligibility tests when tested several times? Test/retest *intra*subjective reliability of intelligibility was assessed by Logan et al. (1989) and Van Bezooijen (1988); in both cases it was found to be good. More attention has been paid to *subject dimensions* systematically affecting *inter*subjective reliability. This research was motivated by the finding of large variance in the test scores, possibly obscuring effects of the synthesis systems compared. Most studies in this area examined variability in intelligibility scores. Subject dimensions considered relevant include: age, non-expert experience with synthetic speech, expert experience with synthetic speech, and analytic listening.

Within the ESPRIT-SAM project (Howard-Jones 1992a,b), the effect of *age* was examined with Italian VCV-items. Five age categories were distinguished (10–19, 20–29, 30–44, 45–59, over 60), with between 5 and 8 subjects per group. The group scores of percentages correct consonant identification ranged from 58%, for the oldest group, to 64%, for the youngest group. So, little evidence was found for an effect of the subject dimension age.

Non-expert experience with synthetic speech was investigated in several studies. Howard-Jones (1992a,b) compared the performance of 8 subjects experienced with synthetic speech and 24 inexperienced subjects. German VCV-items were presented. The mean score for the experienced subjects was 79%, that for the inexperienced subjects 62%. There is further evidence that the intelligibility of synthetic speech increases as a result of non-expert experience with synthetic speech, both when acquired in the form of training with feedback (e.g. Greenspan et al. 1985; Schwab et al. 1985) and when acquired in a more natural way without feedback (Pisoni et al. 1985b,a; Boogaart and Silverman 1992). The learning effect has been

found to manifest itself after only a few minutes of exposure. However, there are indications that the effect of learning depends on the type of synthesis used. Jongenburger and Van Bezooijen (1992) assessed the intelligibility of two synthesis systems used by visually handicapped for reading a digital daily newspaper in a first confrontation and after one month of experience. An open response CVC identification test was used. For one system, which was allophone based, consonant intelligibility increased from 58% to 79%; for the other system, which was diphone based, intelligibility increased from 63% to 68%. It was hypothesised that the characteristics of allophone-based synthesis are easier to learn because they are rule-governed and therefore more invariant than those of diphone-based synthesis. Moreover, no transfer was found from experience with one type of synthesis to the understanding of the other type of synthesis. This suggests that there is no such thing as general experience in listening to synthetic speech.

The subject dimension *expert experience with synthetic speech* was examined by Howard-Jones (1992a) with English VCV-items. A percentage correct consonant identification of 30% was obtained for the inexperienced subjects versus 49% for the experts. So, again improved performance was found as a function of increased exposure.

The last subject dimension we want to mention is *experience in listening* analytically to speech. On the basis of a reanalysis of the results from a number of their evaluation studies, Van Bezooijen and Pols (1993) conclude that the more ear-training subjects have, the higher the percentages correct they attain. Furthermore, ear-training was found to result in a reduction of intersubjective differences.

Apart from variance which can be attributed to particular subject dimensions, much apparently *individual variability* is found in test scores. Hazan and Shi (1993) examined the variance in subject scores in various tests, including intelligibility of meaningless VCV-items, intelligibility of Semantically Unpredictable Sentences (SUS, see Section 12.7.7), and speech pattern identification for plosive place and voicing contrasts. A homogeneous group of subjects was used:

- All had normal hearing (average pure tone thresholds of 20 dB HL or better from 0.25 to 8 kHz in both ears.
- All were students at University College London.
- All fell within the age range 18 to 32 years.
- All were native speakers of English.
- None had any training in phonetics.
- All had little or no previous experience with synthetic speech.

Despite the homogeneity of the subject group, a sizeable degree of variability was found in all tests. For the SUS the range (i.e the difference between the best and worst performing subject) was 28%, for the CVC-test the range was 47%. At the level of speech pattern processing, considerable differences were found in the perceptual weighting given to individual cues to plosive place and voicing contrasts. Hazan & Shi attribute the variability not to audiological differences among listeners, but to the development of different perceptual strategies during language acquisition. They distinguish two types of listeners: "auditors" (i.e. users of acoustic information) and "comprehenders" (i.e. users of global contextual information).

Having established that there is much variability in the scores obtained in speech output evaluation tests, part of which can be attributed to clearly identifiable subject dimensions such as previous experience with synthetic speech, one may wonder what implications this has for the selection of subjects in specific tests. We think that the implications for subject selection depend in part on, at least, the type of listening required (e.g. global versus analytic mode), and the width of application (general public versus specific user groups). Therefore, for some common applications the following recommendations can be formulated:

12.3.1.1 Recommendations on choice of subjects

- 1. Exclude hearing-impaired subjects from speech output assessment. Within the SAM project (Howard-Jones 1992a,b) it is specified that subjects should pass the hearing screening test at 20 dB HL at all octave frequencies from 500 to 4000 Hz.
- 2. Do not use the same subject more than once.
- 3. In diagnostic testing only include subjects speaking the same language (variety) as the language (variety) tested.
- 4. For diagnostic purposes requiring analytic listening, hire a trained phonetician (with a basic understanding of the relationships between articulation and acoustics) in the initial stages of development of a system in order to obtain subtle information (e.g. degree of voicing in plosives), or information that is usually not used for functional purposes in real-life communication (e.g. formal aspects of temporal organisation and intonation, cf. Terken (1993)).
- 5. In specialised applications, select subjects who are representative of the (prospective) users. For example, synthesis integrated in a reading machine for the blind should be tested with visually handicapped. And synthesis for long-term use should be tested with subjects with different degrees of experience and familiarisation with the type of synthetic speech of interest.

The above recommendation was made not only because of (possible) differences in the perception of the speech output, but also because motivation is known to play an important role in the effort people are willing to spend in order to understand suboptimal speech. If people have a choice between human and synthetic speech, the synthetic speech will have to be good if it wants to have a chance of being accepted. However, if people do not have a choice, e.g. the visually handicapped who without synthesis (or braille)

will not have access to a daily newspaper, synthesis will be accepted more easily.

6. Synthesis to be used by the general public for incidental purposes, i.e. which should be functionally adequate in a first confrontation, should be tested with a wide variety of subjects, including people with a limited command of the language, dialect speakers, and people of different ages. However, none of them should have experience in listening to synthetic speech. In telecommunications research, groups of between 12 and 16 subjects (all with English as their primary language) have been found sufficient to obtain stable mean values in judgment tests.

12.3.2 Test procedures

As indicated in Section 12.2, speech output assessment techniques can be differentiated along a number of parameters, but no parameters related to the actual test procedure were included there. Test procedures can vary with respect to subjects (see Section 12.3.1), stimuli, and response modality. *Stimuli* can vary along a large number of parameters, the most important of which are listed below.

- Length and complexity: (e.g. at the word phonology level: monosyllabic, disyllabic, polysyllabic, including only single consonants and vowels or also sequences of consonants and vowels). The more varied in length and complexity the test items are, the more diagnostic information can be obtained and the more representative the test results are for the perception of unrestricted speech output. However, higher linguistic levels are often less suited for diagnostic purposes because subjects' responses are determined by many other sources of information in addition to the acoustic properties of the stimuli (see Section 12.4.1.1).
- Linguistic level: (word, sentence, paragraph). Again, the higher the linguistic level, the better test results can be generalised to unrestricted speech output.
- Stimulus set: (fixed set, where all items are presented each time the test is run, versus open set, where each time new (combinations of) test items are presented, e.g. the SUS Test in Section 12.7.7). Of course, in the light of learning effects open sets are more useful and flexible than fixed sets.
- Meaningfulness: either at the word level or at the sentence level (meaningful, meaningless, or mixed, i.e. lexically or semantically unpredictable). Each choice seems to have both advantages and disadvantages/restrictions. For example, tests which only use meaningful test items at the word level, such as the DRT and MRT (see Sections 12.7.4 and 12.7.5) have the advantage of being reliable and easy to administer. However, intelligibility may be overestimated, there is a risk of a ceiling effect, and they have little diagnostic value. In principle, the mixed approach seems a good choice, because the subjects are not guided in any way as to what constitutes a legal or an illegal response. Nevertheless, there may be a risk of a bias towards meaningful words. For other implications of the choice between meaningful,

meaningless, and mixed items at the word level, see Section 12.5.2.1.2.1. For implications at the sentence level, see Section 12.5.2.1.2.2.

• Representativeness: e.g. Phonetically Balanced (PB) stimulus lists, with a frequency of occurrence of phonemes in accordance with the phoneme distribution in the language tested or the specific domain of application at hand, or equal representation of each phoneme. If one wants to obtain a global idea of the intelligibility of a system, PB-lists are to be preferred, if one aims at diagnostic information, one usually opts for equal representation.

In Section 12.7, summary descriptions of tests are given where the stimuli have been categorised along these stimulus parameters. Chapter 9 on methodology should also be consulted.

Response modality can vary along a number of parameters as well. The choice seems to be mainly determined by three factors: comparative versus diagnostic, functional versus judgment, and TTS development versus psycholinguistic interest. In the five types of response modalities listed below, 1 and 2 are mainly used within the glass box approach (1 in TTS development, 2 in psycholinguistically oriented research), whereas 3, 4 and 5 are more common in the black box approach. The latter three response modalities can be further differentiated in that 3 and 4 are functional in nature (3 in TTS development, 4 in psycholinguistically oriented research), whereas 5 represents judgment testing. In the list of response modalities a distinction is made between off-line tests, where subjects are given some time to reflect before responding, and on-line tests, where an immediate response is expected from the subjects, tapping the perception process before it is finished.

- OFF-LINE IDENTIFICATION TESTS, where subjects are asked to transcribe the separate elements (sounds, words) making up the test items. This response modality can be further differentiated. With respect to the *nature* of the set of response categories there is a choice between:
 - a *closed set*, where subjects are forced to select the appropriate response from a limited number of pregiven categories, and
 - an *open response mode*, where the only restriction are the constraints imposed by the language.

TRANSCRIPTION can be:

- in normal spelling, leading to problems in the interpretation of the responses in case of meaningless or lexically unpredictable stimuli (e.g. if subjects write down "lead", have they heard /led/ or /li:d/?), or
- *unambiguous notation*, placing the burden upon the subjects, since they have to be trained to systematically apply this notation system.

- ON-LINE IDENTIFICATION TESTS, requiring the subject to decide whether the stimulus does or does not exist as a word in the language (so-called lexical decision task, e.g. Pisoni et al. 1985b,a).
- OFF-LINE COMPREHENSION TESTS, in which content questions have to be answered in an open or closed response mode (e.g. Pisoni et al. 1985b,a).
- ON-LINE COMPREHENSION TESTS, requiring the subject to indicate whether a statement is true or not (so-called sentence verification task, e.g. Manous et al. (1985)).
- JUDGMENT TESTS (also called opinion tests), involving the rating of scales (e.g. Pavlovic et al. 1990; Delogu et al. 1991; ITU-T 1993).

The last response modality will be discussed in some more detail. Pavlovic and co-workers have conducted an extensive series of studies (cf. Pavlovic et al. 1990) comparing different types of scaling methods that can be used in judgment tests to evaluate speech output. Much attention was paid to:

- the magnitude estimation method, where the subject is presented with an auditory stimulus and is asked to express the perceived strength/quality of the relevant attribute (e.g. intelligibility) numerically ("type in a value") or graphically ("draw a line on the computer screen"), and
- the *categorical estimation method*, where the subject has to select a value from a limited range of prespecified values, e.g. 1 representing extremely poor and 10 excellent intelligibility.

Pavlovic et al. stress that there are important differences between the two types of scaling methods, for example the fact that categorical estimation results in an interval scale, whereas magnitude estimation results in a ratioscale. The former leads to the use of raw ratings, the calculation of the arithmetic mean, and the comparison of conditions in terms of differences, the latter leads to the use of the logarithm of the ratings, the geometric mean, and comparison in terms of ratios. The differences also have implications for the type of conclusions to be drawn from the test results. Both the categorical estimation method (with a 20-point scale) and the magnitude estimation method have been included in SOAP as standard SAM Overall Quality test procedures (see Section 12.7.11).

12.3.2.1 Recommendations on choice of response modality

- 1. For rapid judgment testing, use intra-subject ("internal comparison") categorical estimation,, and when you do, use at least a 10-point scale.
- 2. To compare results across tests ("external comparison"), use magnitude estimation and when you do, use the line length drawing procedure, asking subjects to express the quality of the stimulus relative to the most ideal (human) speech they can imagine.

12.3.3 Benchmarks

By a *benchmark test* we mean an efficient, easily administered test, or set of tests, that can be used to express the performance of a speech output system (or some module thereof) in numerical terms. The *benchmark* itself is the value that characterises some reference system, against which a newly developed system is (implicitly) set off. The benchmark is preferably chosen such that it represents a performance level that is known to guarantee user satisfaction. Consequently, if the performance of a new product exceeds the benchmark, its designer or prospective buyer is assured of at least a satisfactory product, and probably even better. Obviously, testing against a benchmark is more efficient than pairwise or multiple testing of competing products.

At this time it is too early to talk about either existing benchmarks or benchmark tests. It is clear, however, that the development of benchmarking deserves high priority in the speech output assessment field. As a first step, existing tests should be scrutinised for their potential use as benchmark tests. Choices should be made as to what aspects to include in benchmark tests (overall performance, composite performance by a number of crucial modules), and what system to adopt as the reference on which the benchmark value should be based. In this respect, it seems to us that one should not adopt the performance of human speech as the benchmark. Human speech, at least when produced by professional talkers, will simply be too good for the purpose of benchmarking. Since human speech will always be superior to synthetic speech, the quality of the latter will have to be expressed as a fraction, which makes it hard to compare the relative differences between different types of synthetic speech. What we need is a speech output system of proven, but still imperfect, quality. This is, quite probably, the reason why the quality of many speech output systems for English is often expressed relative to the "Paul" voice of MITalk/DECTalk, which has long served as the *de facto* standard in TTS.

12.3.4 Reference conditions

Next to a widely accepted benchmark, it would appear to us that designers of speech output systems should want to know how well their systems perform relative to some optimum, and what performance could be expected of a system that contains no intelligence at all. In other words, the designer is looking for *topline* and *baseline* reference conditions. *Reference conditions* such as these do not yield diagnostic information in the strict sense of the word. However, they do provide the systems developer with an estimate of how much improvement can still be made to a system as a whole (in a black box approach) or to specific modules (in a glass box approach).

12.3.4.1 Segmental reference conditions

There has been no general practice to include topline and baseline reference conditions in segmental quality testing (Section 12.5.2.1). Still, it seems to us that it is important to reach consensus on a number of measures. If the output system uses waveform concatenation techniques, the designer will want to know how well the synthesis performs relative to the live human speaker, or to facilitate procedures, to some electronic operationalisation of live speech (e.g. CD quality speech recorded at a short distance from the speaker's mouth in a quiet environment). However, if the system's waveforms have been coded with a lower bitrate than CD quality, the designer should determine to what extent degradation of system performance is due to the synthesis itself as opposed to the non-optimal bitrate. An easy way to determine this, is to adopt a second reference condition using the same (lower) bitrate as the synthesis. This precaution is even more necessary for parametric synthesis. Obviously, no type of parametric synthesis can be better than the maximum quality that is afforded by the analysis-resynthesis coding scheme adopted for the synthesiser. This requirement can generally be fulfilled when LPC synthesis schemes are used. However, for a range of synthesisers (e.g. the Klatt and the JSRU synthesisers) no automatic parameter estimation for straightforward analysis-resynthesis is possible at this time. The optimal parametric representation of human reference materials will then have to be found by trial and error (i.e., by adjusting parameter values while making auditory/spectrographic comparisons between the synthesis and the human original), or else the attempt should be abandoned. The designer of an output system claims that the intelligence incorporated into the synthesis systems (e.g. through rules) makes the systems perform better than with no intelligence built in at all. In order to establish the extent to which this claim is true, a baseline condition is needed which consists in a type of synthetic speech that has no knowledge of speech processes at all.

12.3.4.2 Recommendations on choice of segmental reference conditions

- 1. Absolute segmental topline: In the case of allophone synthesis, use human speech produced by a designated talker, i.e. the same individual on whose speech the table values and synthesis rules were based, or who, in the case of concatenative synthesis, provided the basic synthesis building blocks. The absolute topline reference will then be based on CD-quality digital speech.
- 2. Relative segmental topline for parametric synthesis: A second useful topline reference is the human reference speech but analysed and (re-)synthesised using exactly the same coding scheme that is employed in the speech output system to be tested.
- 3. Relative segmental topline for waveform concatenation: Use the same

(lower) bitrate in the reference condition as in the speech output system.

- 4. Segmental baseline for allophone synthesis: Use speech in which all segments retain their table values and are strung together merely by smoothing spectral discontinuities at segment boundaries.
- 5. Segmental baseline for concatenative synthesis: Use speech made by stringing together coarticulatory neutral phones (i.e. stressed vowels spoken between two /s/-es, or stressed consonants preceded by schwa and followed by an unrounded central vowel, cf. the "neutrone" condition in Van Bezooijen and Pols (1993). Minimal smoothing should be applied to avoid spectral jumps.

12.3.4.3 Prosodic reference conditions

The need for suitable topline and baseline reference conditions has clearly been recognised in the field of *prosody* (i.e. temporal and melodic structure, cf. Section 12.5.2.3) testing. The following are recommendations for prosodic topline and baseline conditions. Note that, in contrast to segmental evaluation, listeners often find it very difficult to differentiate between different prosodic versions of an utterance. Therefore testers often need examples of "very bad" systems to check whether the listeners are indeed sensitive to prosodic differences.

12.3.4.4 Recommendations on choice of temporal reference conditions

- 1. Temporal and melodic topline: Copy, as accurately as possible within the limitations of the synthesiser, the temporal structures and speech melodies of a single designated professional human speaker onto the synthetic speech output.
- 2. Temporal baseline: Use a condition in which the smallest synthesis building blocks (phoneme, diphone, demisyllable) retain their original, unmanipulated durations as they were copied from the human original from which they were extracted (or, in the case of allophone synthesis, the phoneme duration table values (cf. Carlson et al. 1979).

This baseline condition, then, contains no intelligence, so that any improvement in the target conditions with duration rules must be due to the added explicit knowledge on duration structure. A reference in which segment durations vary at random (within realistic bounds) can be included for validation purposes, as an example of a "very bad system". Listeners should rate this condition as poorer than any other condition.

12.3.4.5 Recommendations on choice of melodic reference conditions

1. Melodic baselines: Synthesise utterances on a monotone, at a pitch level that coincides with the average pitch of the test items. Also, include a random melodic reference for the sake of validation, by introducing random pitch variations (in terms of excursion size, rate of change, and segmental

alignment), within physiologically and linguistically reasonable limits and with a mean pitch equal to the average of the test items.

There is a practical problem that not every synthesiser allows the generation of monotonous pitch so that some sort of waveform manipulation (e.g. pitch synchronous overlap and add, PSOLA) may have to be used in order to monotonise the synthetic melody.

12.3.4.6 Voice characteristics reference conditions

In the area of *voice characteristics* (voice quality, Section 12.5.2.4), the problem of reference conditions has not been recognised. Generally, there seems to be little point in laying down a baseline reference for voice quality. The choice of a suitable topline would depend on the application of the speech output system. If the goal is personalised speech output (for the vocally handicapped) or automatic speaker conversion (as in interpreting telephony), the obvious topline is the speaker who is being modelled by the system, using the same coding scheme when applicable. When a general purpose (i.e. non-personalised) speech output system is the goal, one would first need to know the desired voice quality, i.e. ideal voices should be defined for specific applications, and speakers should be located who adequately represent the ideal voices. At this time we will refrain from making any further suggestions on this matter. The definition of "ideal" voices and voice qualities, and the implementation of topline references should be a matter of priority in the near future.

12.3.4.7 Overall quality reference conditions

Given the existence of an overall quality topline reference condition, it would be advantageous to have a set of reference conditions that are poorer than the optimum by a number of calibrated steps until a quality equal to or less than the baseline reference is reached (see also Section 12.4.1.3). Such a set of reference conditions would yield a grid within which each type of speech, whether produced by humans or by machines, can be located and compared with other types of speech. Recently, attempts have been made at creating such a continuum of reference conditions by taking highquality human speech and applying some calibrated distortion to it, such as multiplicative white noise at various signal-to-noise ratios ("Modulated Noise Reference Unit or MNRU", cf. ITU-T Recommendation P.81), or time-frequency warping (TFW, ITU-T Recommendation P.85 (cf. Burrell 1991), or T-reference (cf. Cartier et al. 1992).

TFW introduces greater or lesser (random) deviations from the mean rate of a recorded utterance ($\pm 2.5\%, \pm 5\%, \ldots, \pm 20\%$) over successive stretches of 150 ms, so that the speech contains potentially disturbing accelerations and decelerations and associated frequency shifts. Fellbaum et al. (1994) showed that the MNRU is not suitable for the evaluation of synthetic speech. TFW of natural speech, however, provided a highly sensitive reference grid within which TTS systems could be clearly differentiated from each other in terms of judged listening effort (Johnston 1993). Moreover, Johnston showed that the perceived quality ordering among a range of TTS systems interacts with the sound pressure level at which the speech output is presented.

12.3.4.8 Recommendations on choice of overall quality reference conditions

1. Use time-frequency warping of optimal human speech to create a grid of overall quality reference conditions.

12.3.5 Comparability across languages

Although it is generally agreed that in the final analysis all languages are equally complex, it cannot be denied that phonetic and phonological complexity differs widely from one language to the next. Languages differ in the size of their vowel and consonant inventories, in the complexity of syllable structures, stress rules, reduction processes, and so forth.

A number of systems are (commercially) available that provide multilingual speech output (e.g. DECTalk, INFOVOX, MULTIVOX, APOLLO). Generally, such systems were primarily developed for one language (American English, Swedish, Hungarian, and British English, respectively), and additional language modules were derived from the original language by making minimal changes to the basic units and rules. As a result it is commonly observed that the derivate languages of multilingual systems sound poorer than the original. Yet, it is very difficult to establish this convincingly, since the poorer performance may be due (completely or in part) to the greater intrinsic difficulty of the sound system of the new language. Ultimately one would like to develop speech output assessment techniques that allow us to determine the quality of a system speaking language A and to compare its quality to that of another system speaking language B. In order to reach this objective, we would have to know how to weigh the scores obtained for a language for the intrinsic difficulty or complexity of the relevant aspects in that language.

Such goals will not easily be accomplished. However, steps have been taken in the SAM project to ensure optimal cross-language comparability in the construction of the test materials and administration procedures. For example, in the Semantically Unpredictable Sentence Test (SUS Test, see Section 12.7.7), the same five syntactic structures (defined as linear sequences of functional parts of speech, e.g. Subject–Verb–Direct Object) are used in all languages tested, and words are substituted in each of the designated syntactic slots that are selected from the same lexical categories, and with the shortest word length allowed by the language (see Section 12.7.7). It should be obvious, however, that complete congruence cannot be obtained in

this fashion: the shortest content words in Italian and Spanish are typically disyllables, while they are monosyllabic in French and the Germanic languages. Similarly, although all five syntactic structures occur in each of the languages tested, certain structures will be more common in one language than in an other. Given the existence of such intrinsic and unavoidable structural differences between languages, we recommend further research into the development of valid cross-language normalisation measures.

Especially when working within the European Union, with its increasing number of partner countries and languages, speech output products are likely to be produced on a multilingual basis. The further development of efficient testing procedures that can be validly used for all relevant languages is a clear priority. Yet, as explained above, we should not raise our hopes too high in this matter, given the existence of intrinsic and unavoidable structural differences between languages. For this reason we recommend parallel research into the development of valid cross-language normalisation measures that will allow us to realistically compare speech output test results across languages, if the choice of test materials cannot be balanced in all relevant linguistic aspects.

In this effort, ITU recommendation P.85 has potential. Following this procedure (see Section 12.3.4.7) a reference grid can be constructed for each (EU) language. One possible outcome could be that some languages prove more resistent to time-frequency warping than others, although we hesitate to make any predictions. Be this as it may, differences in intelligibility between languages would be effectively normalised out when we determine the quality of an output system relative to the reference grid that is applicable for the language being tested.

12.4 Black box approach

12.4.1 Laboratory testing

12.4.1.1 Functional laboratory tests

Black box assessment tests a system's performance as a whole, without considering the performance of modules internal to the system. Ideally, within black box testing, one would want to have at one's disposal a functional test to assess the adequacy of the complete speech output in all respects: does the output function as it should? Such a test does not exist, and is difficult to conceive. In practice, the functional quality of overall speech output has often been equated with comprehensibility: to what extent can synthesised continuous speech be understood by listeners?

Speech comprehension is a complex process involving the interpretation and integration of many sources of information. Important sources of information in complete communication situations, where both auditory and visual information are available to interactants, are:

- 1. Speech signal information at different levels (segments, prosody, voice characteristics),
- 2. Segment combinatory probabilities (e.g. /str../ is a permissible consonant sequence at the onset of words in many EU languages, but all other permutations of this sequence are illegal (e.g. /tsr.../,
- 3. Knowledge of which segment strings are existing words in the language (e.g. the permissible string /strik/ is not a word in English),
- 4. Word combinatory probabilities (e.g. the article "the" will tend to be followed by adjectives or nouns rather than verbs),
- 5. Semantic coherence (e.g. in the context of "arrive" a word like "train" is more probable than a word like "pain"),
- 6. Meaning extracted from the preceding linguistic context; due to the repetition of words and the progressive building up of meaning, the last sentence of a text will generally be easier to understand than the first,
- 7. World knowledge and expectations of the listener based on previous experience,
- 8. Cues provided by the extra-linguistic context in which the message is spoken (e.g. facial expressions and gestures of the speaker, relevant things happening in the immediate environment).

In normal daily life all these different sources, and others, may be combined by listeners to construct the meaning of a spoken message. As a result, in applied contexts the contributions of separate sources are difficult to assess. Laboratory tests typically try to minimise or control for the effects of at least some of the sources in order to focus on the auditory input. Some segmental intelligibility tests at the word level (such as the SAM Standard Segmental Test, see Section 12.7.1) try to minimise the effects of all sources except (1) and (2): only meaningless but permissible consonant–vowel–consonant combinations (e.g. /hos/) or even shorter items (/ze, ok/) are presented to the listener. In comprehensibility tests, factor (8) is excluded completely and (7) as far as possible. The latter is done by selecting texts with supposedly novel information for all subjects.

No completely developed standardised test, with fixed test material and fixed response categories, is available for evaluating comprehension, but one wonders whether this would be very useful in the first place, since it is not clear what the "average" text to be used should look like in terms, for example, of the complexity and type of vocabulary, grammatical structures, sentence length, and style. At this level of evaluation it advisable to take the characteristics of the intended application into account.³

Testing the comprehensibility of speech output destined to provide traffic information requires a more specific type of test materials (e.g. short sentences, only statements, restricted range of lexical items, formal style) than speech output to be used for reading a digital daily newspaper for the blind,

 $^{^{3}}$ Clearly, there is a continuum from completely application independent at the one end and completely application specific at the other end.

where the test materials should be more varied in all respects. The greatest variation should probably be present in speech material testing text-tospeech systems developed to read novels to the visually handicapped.

As to the type of comprehension test, several general approaches can be outlined. The most obvious one involves the presentation of synthesised texts at the paragraph level, preferably with human produced versions as a topline control, with a series of open or closed (multiple choice) questions. Results are expressed in terms of the percentage of correct responses. An example of a closed response approach is Pisoni et al. (1985a,b), who used 15 narrative passages selected from standardised adult reading comprehension tests. Performance was compared between listening to synthetic speech, listening to human speech, and silent reading. Each condition was tested with 20 subjects. One of the most important findings was a strong learning effect for synthetic speech within a very short time, and the absence of clear differences among the test conditions.

At first sight, the results of closed response comprehension tests seem counterintuitive: although the human produced texts sound better than the synthetic version, often no difference in comprehension is revealed (Nye et al. 1975; Delogu et al. 1992b) or, after a short period of familiarisation, even superior performance for synthetic speech (Pisoni et al. 1985b,a) is observed. These results have been tentatively explained by hypothesising that subjects may make more of an effort to understand synthetic speech. This could be expected to lead to:

- Slower reaction times in a sentence verification test.
- A decrease in performance as a function of fatigue.
- Poorer performance for secondary tasks.

Confirmation of the first prediction was found by Manous et al. (1985). The second and third predictions were tested by Luce et al. (1983), using a word recall test, and by Boogaart and Silverman (1992), using a tracking task. The first study revealed a significant effect, whereas the second did not. However, the lack of differentiation in comprehensibility between human and synthetic speech in the above studies may also be due to the use of the closed response approach, where subjects have a fair chance of guessing the correct answer. Open response tests are known to be more sensitive, i.e. more apt to bring to light differences among test conditions. An example of an open response study is Van Bezooijen (1989), who presented five types of texts typically found in daily Dutch newspapers, pertaining to the weather, nature, disasters, small events, and sports, to 16 visually handicapped subjects. An example of a question testing the comprehensibility of the weather forecasts is: What will the temperature be tomorrow? The questions were sensitive enough to yield significant differences in comprehensibility among two text-to-speech (one automated and one manually

corrected) and one human produced version of the texts. Crucially, the results also suggest that the effect of the supposedly greater effort expended in understanding synthetic speech has its limits. If the synthetic speech is bad enough, increased effort cannot compensate for loss of quality.

The tests described ask subjects to answer questions after the texts have been presented, thus measuring the final product of text interpretation. In addition to these off-line tests, more psycholinguistically oriented online approaches have been developed which request instantaneous reactions to the auditory material being presented. These tests primarily aim at gaining insight into the cognitive processes underlying comprehension: to what extent is synthetic speech processed differently from human speech? A few of these psycholinguistic tests are:

- The word monitoring task, requiring subjects to press a button as soon as they hear a prespecified word.
- The *sentence-by-sentence listening task*, in which subjects push a button whenever they are ready to listen to the next sentence (comprehension is checked afterwards but is not part of the test proper).
- The sentence verification test, where subjects have to decide whether short sentences are true statements or not (e.g. Mud is dirty and Rockets move slowly).

All three are on-line measures, the first indexing cognitive workload, the second and third assessing speed of comprehension. On-line tests of this type, which invariably reveal differences between human and synthetic speech, have been hypothesised to be more sensitive than off-line measures (Ralston et al. 1991). However, the results of such psycholinguistic tests ("subjects responded significantly faster to system A (740 ms) than to system B (930 ms)") are less interpretable for non-scientists than those of comprehension tests ("subjects answered 74% of the system A questions correctly versus 93% of the system B questions"). On the other hand, insight into cognitive load may ultimately prove important in dual task applications.

12.4.1.2 Recommendations on functional testing of overall output quality

- 1. Try to avoid the use of functional tests to assess overall output quality: online reaction time tests are difficult to interpret and off-line comprehension tests are difficult to develop.
- 2. If determined to develop a comprehension test, beware of the fact that reading tests may be too compact to be used as listening tests; adapt the materials or use materials that are meant to be listened to.
- 3. Use open comprehension questions rather than closed ones, the former being more sensitive than the latter.
- 4. When administering a comprehension test, include a topline reference with a dedicated speaker realising exactly the same texts presented in a synthetic version; use different groups of subjects for the various speech con-

ditions (or better still block conditions over listeners such that no listener hears more than one version of the same text while at the same time each listener gets an equal number of different text versions).

5. When interpreting comprehension results, look at difference scores (synthetic compared to human) rather than at absolute scores in order to abstract from the intrinsic difficulty of questions.

12.4.1.3 Judgment laboratory tests

The black box tests described so far are functional in nature. However, instead of evaluating overall quality functionally, subjects can also indicate their subjective impression of global quality aspects of synthetic output by means of rating scales. Taking comprehensibility as an example, a functional task would be one where subjects answer a number of questions related to the content of a text passage as described above. Alternatives from a judgment point of view include:

- *Paired comparison*, where subjects indicate which of two synthesisers sounds more comprehensible,
- *Magnitude estimation*, where subjects assign a value expressing, or draw a line of a length which is equal to the magnitude of, their impression of comprehensibility,
- *Categorical estimation*, where subjects rate synthesisers, for instance, along a 10-point scale which runs from 1: extremely incomprehensible to 10: extremely comprehensible.

Some methodological aspects of the second and third method are described in detail in Section 12.3.2. There it is also indicated that magnitude estimation is relatively laborious and better suited to test external comparison, whereas categorical estimation is relatively fast and easy, and better suited to test internal comparison.

Both the magnitude (continuous scale) and categorical estimation (20-point scale) methods have been included in SOAP in the form of the SAM Overall Quality Test (see Section 12.7.11). Three judgment scales are recommended, related to:

- Intelligibility (How identifiable does the message sound?),
- *Naturalness* (To what extent does the message sound like being produced by a human speaker?),
- Acceptability (The overall user's satisfaction with the communicative situation).

The intelligibility and naturalness ratings are based on pairs of (unrelated) sentences. Fixed lists of 160 sentences of varying content and length are available for Dutch, English, French, German, Italian, and Swedish. Examples for English are: *I realise you're having supply problems but this is*

rather excessive and I need to arrive by 10.30 a.m. on Saturday. For the acceptability ratings, application specific test materials are recommended. The magnitude and categorical estimation procedures have been applied to speech output in a number of studies (e.g. Pavlovic et al. 1990; Delogu et al. 1991; Goldstein et al. 1992). Methodological aspects, such as the effects of stimulus range and the number of categories, relationships among methods, reliability, and validity, are emphasised.

The importance of application-specific test materials is also stressed by the International Telecommunication Union Telecommunication Standardisation (ITU-T) sector (see Section 12.7.12). They developed a test specifically aimed at evaluating the quality of telephone speech (where synthesis could be the input). It is a categorical estimation judgment test comprising ratings on (a subset of) eight scales:

- Acceptance
- Overall impression
- Listening effort
- Comprehension problems
- Articulation
- Pronunciation
- Speaking rate
- Voice pleasantness

The first scale is a 2-point scale, the others are 5-point scales. Strictly speaking, only the first four scales can be captured under the heading overall quality; the other four scales are directed at more specific aspects of the output and require analytic listening. The content of the speech samples presented should be in accordance with the application. Examples of application-specific test items are: Miss Robert, the running shoes Adidas Edberg Pro Club, colour: white, size: 11, reference: 501-97-52, price 319 francs, will be delivered to you in 3 weeks (mail order shopping) and The train number 9783 from Poitiers will arrive at 9:24, platform number 3, track G (railway traffic information). In addition to rating the eight scales, subjects are required to reproduce information contained in the message. A pilot study has been run by Cartier et al. (1992). Fellbaum et al. (1994) tested 13 synthesis systems for German using the ITU-T Overall Quality Test as well as open response functional intelligibility tests. Waveform concatenative synthesis systems proved measurably better than formant synthesis systems.

Van Bezooijen and Jongenburger (1993) employed a similar series of judgment scales as proposed by the ITU-T in a mixed laboratory/field study which addressed the suitability of synthetic speech within the context of a digital daily newspaper for the blind (see Section 12.4.2.2). Their battery comprised ten 10-point scales:

• Intelligibility

- General quality
- $\bullet \ {\rm Naturalness}$
- Precision of articulation
- Accuracy of pronunciation
- Pleasantness of voice
- Adequacy of word stress
- Appropriateness of tempo
- Liveliness
- Fluency

Again a distinction can be made between scales relating to overall quality (the first three scales), and the other scales, relating to specific aspects of the speech output. A factor analysis yielded two factors, the first with high loadings of intelligibility, general quality, and precision of articulation, the second with high loadings of naturalness, pleasantness of voice, and adequacy of word stress. Intelligibility and naturalness were taken by the authors to be the two central dimensions underlying the evaluative judgments.

12.4.1.4 Recommendations on judgment testing of overall output quality

- 1. Since there is no consensus on the most appropriate judgment scales to evaluate overall quality, choose between:
 - Intelligibility, naturalness, and acceptability (SAM Overall Quality Test),
 - Acceptance, overall impression, listening effort, and comprehension problems (ITU-T), or only listening effort (practice in telephony),
 - Intelligibility, general quality, and naturalness (Van Bezooijen and Jongenburger 1993).
- 2. It is important that the scale positions have a clear meaning for the subjects and that the scale is wide enough to allow differentiation among systems compared. Use at least a 10-point scale.

12.4.2 Field testing

12.4.2.1 Preliminary remarks

In the previous section, the black box approach to speech output evaluation was operationalised within a laboratory context. From an experimental point of view, the main advantage of a laboratory study is control over possibly interfering factors. However, ultimately it is the functioning of a speech output system in real life, with all its variability, that counts. If overall quality is extended to include all aspects of the synthesis in the context of an application, testing may be necessary in the field. Due to the variety of applications, it is difficult to summarise the aspects which field tests have in common. To illustrate the diversity, some examples will be given below.

12.4.2.2 Field tests

A combined laboratory/field functional/judgment test, with equal attention to the speech output itself and the context within it is used, was done by Van Bezooijen and Jongenburger (1993). They used the following suite of four tests to evaluate the functioning of an electronic newspaper for the visually handicapped:

- An interview enquiring after the subjects' attitudes towards the technology,
- A functional open response identification test with CVC-words,
- Judgments on 10 evaluative scales for continuous text passages (1: extremely bad, 10: extremely good) related to global quality and more specific aspects of the speech output, such as pleasantness of voice (see Section 12.3.4.6), adequacy of word stress, appropriateness of tempo, liveliness, and fluency (see Section 12.4.1.3 for further details),
- A functional test assessing the subjects' proficiency in finding their way through the newspaper. This test involved a number of searches, such as *Is there an article on Japan in the economy section?* Proficiency was assessed both in terms of the percentage of correct answers and task completion time.

Each of 24 visually handicapped subjects was visited at home, at three points in time. Since the subjects lived scattered all over the Netherlands, administration of the suite of tests was very time consuming.

Comparable studies have been conducted to evaluate a digital daily newspaper in Sweden (Hjelmquist et al. 1987). However, the experimental set-up to assess the quality of various aspects of the Swedish speech output was less strict: most information was obtained through interviews. On the other hand, much emphasis was placed upon the reading habits of the users: all keystrokes were registered during long periods of time, so that the frequency of use of all reading commands (e.g. *next sentence, previous sentence, talk faster, talk in letters*) could be determined.

A semi-field study combining function and judgment testing within the context of telephone information services was done by Roelofs (1987). In this test resynthesised human speech was used, but the set-up and results can be generalised to synthetic speech output. Two applications were considered, namely directory assistance (the subject puts his request to an operator and then the number is spoken by the computer twice, thus freeing the operator for the next subscriber) and a service for train departure times (in a single pre-stored message the departure times of a number of trains with different destinations are given). In the former application a human operator served as a reference, in the second high-quality PCM speech was

presented. Subjects were sent the instructions in advance and dialled the two services from their homes. The availability of interrupt facilities and speaking rate were examined. Both actual performance (success in writing down the requested data), subjective reactions were registered (14 5-point scales such as bad-good, impersonal-personal, inefficient-efficient) and two questions were added, namely: Do you find this way of information presentation acceptable? and Do you think this service could replace the current service? Due to several factors, the results are of limited value. However, the method is a nice example of how different approaches to testing can be combined.

With a view to exploring the possibilities of synthetic speech for a name and address telephone service, Delogu et al. (1993b) tested six Italian TTS systems by presenting lexically unpredictable VCV and CV sequences in an open response format. Intelligibility scores dropped from 31 to 21% when the same materials were listened to through a telephone line rather than good quality headphones. Curiously enough, the best TTS systems suffered most from telephone bandwidth limitation.

Finally, an important area of speech output evaluation are applications where people are required to process auditory information while simultaneously performing some other task, involving hands and eyes, for instance to write down a telephone number or land an aircraft. The requirements imposed by double tasks like these have been simulated for instance by having subjects answer simple questions related to the content of short synthesised messages while at the same time tracking a randomly moving square on a video monitor by moving a mouse (Boogaart and Silverman 1992). This type of laboratory study could and should be extended to more real-life situations. Other important areas are field tests where the functioning of speech output is tested under various noise conditions, and combinations of noise and secondary tasks.

Since field tests will often have to meet specific requirements, it is not realistic to think in terms of standard tests and standard recommendations. Each case will have to be examined in its own right. In order to get an overview of complex test situations that may arise, Jekosch and Pols (1994) recommend a "feature" analysis to define a test set-up, where features are all aspects relevant to the choice of the test. Their analysis comprises three steps, naturally leading to a fourth step:

- 1. Determine the application conditions (What is to be tested? What are the properties of the material generated?), resulting in a feature profile of the application scenario.
- 2. Define the best possible test matching this feature profile.
- 3. Make a comparison of what is desired and what is available in terms of tests.
- 4. Adapt tests or develop your own test.

Because of the specific nature of some applications, often there will be no ready-made test available, so that it is perhaps better to talk of (suggestions for) test approaches than tests. Moreover, a single test will generally not suffice, but a suite of tests will be needed instead. In this suite both functional and judgment tests can be included. Interviews can be part of the evaluation as well. Moreover, it is possible to administer laboratory type experiments in a field situation. This can be done, for example, by preparing stimulus tapes beforehand and playing them to subjects in the environment where the synthesis system will be used.

12.5 Glass box approach

12.5.1 Linguistic aspects

In this section we shall deal with evaluation procedures that have been, or can be, followed when modules in a text-to-speech system yield some intermediary symbolic output. As was stated above, there are no established methods for evaluating the quality of linguistic modules in speech output testing. As a result there is no agreed-upon methodology in this area nor are there evaluation experts; what little evaluation work is done, is done by the same researchers who developed the modules. In view of the lack of an established methodology we will refrain from making recommendations on the use of specific linguistic tests and test procedures. The need for a more general research effort towards a general methodology in the field of linguistic testing will be discussed in Section 12.6.3.

12.5.1.1 Pre-processing

The first stage of a linguistic interface makes decisions on what to do with punctuation marks and other non-alphabetic textual symbols (e.g. parentheses), and expands abbreviations, acronyms, numbers, special symbols, etc. to full-blown orthographic strings, as follows:

abbreviations	"i.e."	\rightarrow	that is
	"viz."	\rightarrow	namely
acronyms	"NATO"	\rightarrow	nay to e
	"UN"	\rightarrow	$you \ en$
$\operatorname{numbers}$	``124"	\rightarrow	one hundred and twenty four
	``1:24"	\rightarrow	twenty four minutes past one
special symbols	``#1"	\rightarrow	number one
	" $\pounds 1.50$ "	\rightarrow	one pound fifty

There are no standardised tests for determining the adequacy of text preprocessors. Yet is seems that all preprocessors meet with the same classes of transduction problems, so that it would make sense to set up a multilingual benchmark for preprocessing. Laver et al. (1988, 1989), describing
the internal structure of the CSTR text preprocessor, mention a number of transduction problems and present some quantification of their errors in the various categories, which we recapitulate in Table 12.1. The test was run on a set of anomaly-rich texts taken from newspapers and technical journals.

Table 12.1: Percentage of correct treatment of textual anomalies by CSTR text preprocessor (after Laver et al. 1988, pp. 12– 15)

Anomaly type	# tested	# correct	% correct
Meta-textuals	95	95	100
Capital-initials	87	87	100
Digit-bearing	35	34	97
Hyphens	15	15	100
Other	24	24	100
Proper names			73
Strings in dictionary			85
Strings not in dictionary			87
Number strings			100

The results in Table 12.1 are not so much revealing in terms of the numerical information they offer as in the taxonomy of errors opted for. The only other formal evaluation of a text preprocessor that we have managed to locate uses a completely different set of error categories. Van Holsteijn (1993) presents an account of a text preprocessor for Dutch, and gives the results of a comprehensive evaluation of the module. It was observed that the use of abbreviations, acronyms and symbols differs strongly from text to text. Three types of newspaper text were broadly distinguished:

- 1. editorial text on home and foreign news,
- 2. editorial text on sports and business,
- telegraphic-style text (i.e. classified adds, film & theatre listings, radio & television guide).

Text segmentation errors were separately counted for:

- Sentence demarcation.
- Expression demarcation (within sentence subunits that need some specific type of preprocessing).

Correctly demarcated expressions could then be characterised further in terms of:

- Labelling errors,
- Expansion errors.

Finally, a distinction is made between unavoidable and avoidable errors. The former type would be the result of incorrect or unavailable syntactic/semantic information that would be needed in order to choose between alternative solutions. The latter type is the kind of error that needs correction, either by the addition of new rules or by inclusion in the exceptions lexicon. Table 12.2 presents some results.

	(1)	(2)
	segm. sent	label. express.
Α.	0.1 (0.4) N=786	$< 0.01 \ (< 0.01) \ N = 13699$
В.	$0.2 (0.4) \text{ N}{=}643$	$< 0.01 \ (< 0.01) \ N = 10388$
С.	$0.0 (0.5) \text{ N}{=}202$	0.0 (2.0) N = 2570
	(3)	(4)
	label. express.	expan. expr.
Α.	0.9 (3.8) N=1904	1.5 (0.0) N = 479
В.	0.3 (4.0) N=1683	0.0 (0.0) N=571
С.	0.9 (2.5) N=1231	0.7 (0.4) N=560

Table 12.2: Evaluation results for text preprocessor TextScan (after Van Holsteijn 1993).

Percentage of avoidable errors in four categories; percentage of unavoidable errors in parentheses; N specifies the 100% base per cell.

The proposals by Laver et al. (1988) and Van Holsteijn (1993) represent rather crude, and disparate, approaches towards a taxonomy of errors of a text preprocessor. What is clearly needed for the evaluation of text preprocessors, is a more principled analysis of the various tasks a text preprocessor has to perform, focussing on those classes of difficulties that crop up in the European language concerned. Procedures should be devised that automatically extract representative items from large collections of recent text (newspapers) in each of the relevant error categories, so that multilingual tests can be set up efficiently. Once the test materials have been selected, the correct solutions to, for instance, expansion problems can be extracted from existing databases, or when missing there, will have to be entered manually.

12.5.1.2 Grapheme-phoneme conversion

By grapheme-phoneme conversion we mean a process that accepts a fullblown orthographic input (i.e. the output of a preprocessor), and outputs a string of phonemes. The output string does not yet contain (word) stress marks, (sentence) accent positions, and boundaries. The correct phonemic representation of a normally spelled word depends on its linear context

and hierarchical position (e.g. assimilation to adjacent words: I have to go /aIhæf təgəU/ but I have two goals /aIhæv tu: gəUlz/; or the choice between heterophonous homographs: I lead /li:d/ but made of lead /lɛd/ (see also Chapter 6). Therefore the adequacy of grapheme-phoneme conversion modules should not, in principle, be tested on the basis of isolated word pronunciation (citation forms). In practice, however, this is precisely what is done. The reasons for this are threefold:

- For many languages pronunciation databases (or machine readable pronouncing dictionaries) are available, which are exclusively based on isolated words.
- The adaptation rules for word pronunciation in context are (believed to be) straightforward, exceptionless, and easy to implement, and
- Machine readable phonemic transcriptions of continuous prose are scarce.

Table 12.3 presents results of a multilingual evaluation of graphemephoneme converters for seven EU languages, performed within ESPRIT 291/860 "Linguistic analyses of European languages," based on isolated word pronunciation. Since it has often been reported that many more conversion errors occur in proper names than in ordinary words, the evaluation distinguished between four types of materials:

- Newspaper texts of at least $100,000~{\rm words},$

Spanish

• The names of the 150 largest towns per country,

98.9

- Names of 31 European capitals in the national spelling, and
- The 150 most frequent Christian names for each language.

19.	91, p. 394)			
Language	Newspaper	Towns	Capitals	First names
Dutch	98.9	85.3	96.7	89.4
English	90.3	46.0	58.0	58.0
French	96.9 - 94.5	77.3	74.2	84.7
German	93.0 - 90.0	81.0	61.0	80.0
Greek	98.7	97.3	93.5	97.3
Italian	85.2	85.3	80.6	86.9

Table 12.3: Percentage correct grapheme-phoneme conversion in seven EU languages in four types of materials (after Pols 1991, p. 394)

Note: Newspaper scores are weighed for token frequency. Higher first score for French excludes all preprocessing errors; higher first German score is based on the use of an exceptions list.

95.3

96.6

98.0

Incidentally, the results should not be taken to indicate that spelling is harder to convert to phonemes in Italian than in any other language, since different conversion methods were used for each language; however, Italian proper names are no more a problem than ordinary text words. In English and French spelling the proper names do present a serious problem, so that exceptions lists will be a priority for these languages.

In a complementary test Nunn and Van Heuven (1993) compared the performance of three grapheme-phoneme converters for Dutch, i.e. two systems with no or only implicit morphological decomposition (Kerkhoff et al. 1984; Berendsen et al. 1986) and one that included the MORPA morphological decomposition module. About 2,000 simplex and complex (see Section 12.5.1.4) test words were selected from newspaper texts that did not belong to the 10,000 most frequent Dutch words, so that dictionary look-up would fail. Phoneme, syllabification, and stress placement errors were found by automated comparison with a hand-made master transcription file. The earlier converters performed at a success rate of 60% and 64%, which is considerably poorer than the newspaper text score in Table 12.3. The newer system with explicit morphological decomposition was correct in 78%.

12.5.1.3 Word stress

Stressed syllables are generally pronounced with greater duration, greater loudness (in terms of acoustical intensity as well as pre-emphasis on higher frequencies), and greater articulatory precision (no consonant deletions, more peripheral vowel formant values). Moreover, when a word is in focus, a prominence-lending fast pitch movement occurs on the stressed syllable of that word. Except for French, where stress is always on the last full syllable of the word, the stress position varies from word to word in all other EU languages. However, stress position in these languages is predictable to a large extent on the basis of:

- the internal make-up of words (in terms of their constituent morphemes, the hierarchical relationships between adjacent morphemes, and the lexical categories of these morphemes), and
- the segment structure of the syllables making up the morphemes.

All the EU languages have a proportion of idiosyncratic words that do not comply with the proposed stress rules for diverse reasons. Therefore the coverage of stress rule systems has to be evaluated, and errors have to be corrected by including the problematic words in an exceptions dictionary. Tests of stress rule modules have been performed only on an ad hoc basis, either checking the output of the rules by hand (Barber et al. 1989, for Italian), or automatically (using the phonemic transcription field in lexical databases containing stress marks (Langeweg 1988, for Dutch), which

in turn had been checked by hand in some earlier stage of the database development).⁴

12.5.1.4 Morphological decomposition

In morphological decomposition orthographic words are analysed into morphemes, i.e. elements belonging to the finite set of smallest subword parts with an identifiable meaning (see Chapter 6). Morphological decomposition is necessary when the language/spelling allows words to be strung together without intervening spaces or hyphens so as to form an indefinitely large number of complex, longer words. For many EU languages word-internal morpheme boundaries are referred to by the grapheme-phoneme conversion. For instance, if the English letter sequence sh is pronounced as /f/ when it occurs morpheme internally as in *bishop*, but is pronounced as /s/ followed by /h/ when a morpheme boundary intervenes, as in *mishap*.

Obviously, long and complex words will have to be broken up into smaller basic words and affixes (i.e. morphemes) before the parts can be looked up in an exceptions dictionary. If all complex words were to be integrally stored in the lexicon, it would soon grow to unmanageable proportions. For stress placement rules it is sometimes necessary to refer to the hierarchical relationships between the constituent morphemes (e.g. *'lighthouse keeper, light 'housekeeper*, where "'" denotes main stress) and to the lexical category of the word-final morpheme (which generally determines the lexical category of the complex word as a whole, e.g. black+bird is a noun, pitch+black is an adjective). Morphological decomposition is a notoriously difficult task, as one input string can often be analysed in a large number of different ways. The hard problem is choosing the correct solution out of the many possible solutions.

As far as we have been able to ascertain, there are no established test procedures for evaluating the performance of morphological decomposition modules. (Laver et al. 1988, pp. 12–16) tested the morphological decomposition module of the CSTR TTS on 500 words randomly sampled from a 85,000 word type list, which was compiled from a large text corpus as well as from two machine-readable dictionaries. The output of the module was examined by hand, and proved accurate at 70% (which seems rather low considering the fact that the elements of English compounds are generally separated by spaces or hyphens).

The Dutch morphological decomposition module MORPA (MORphological

⁴English presents a special problem in the assignment of stress. The elements of English compounds are typically separated by spaces, so that each element is erroneously treated as a word by itself. Moreover, the stressing of compounds in English partly depends on the semantic relationship between the words that make up the compound, and in part on purely lexical factors. A comparison of English compound stress rules developed by linguists and decision rules automatically extracted from hand-labelled phonetic databases has been reported by Sproat et al. (1992).

PArser, Heemskerk and Van Heuven (1993)) compared the module's output with pre-stored morphological decomposition in a lexical database. In this comparison only segmentation errors were counted, in a sample of 3,077 (simplex and complex) words taken from weekly newspapers. The results showed that in 3% of the input the whole word, or part of it, could not be matched with any entry in the MORPA morpheme lexicon. The frequency of this type of error depends on the coverage of the lexicon. Erroneous analyses were generated in another 1% of the input words. In all other cases the correct morphological segmentation was generated, either as the single correct solution (44%), or as the most likely solution in an ordered list of candidate segmentations (48%), or as one of the less probable candidate solutions (3%).⁵ Although both the accuracy and the coverage of the MORPA module seems excellent by today's standards, the module proved too slow for realistic text-to-speech applications. Processing speed is therefore an important criterion in the evaluation of morphological parsers. There will be a speed/accuracy/coverage trade-off in the evaluation of morphological parsers.

12.5.1.5 Syntactic parsing

Syntactic analysis lays the groundwork for the derivation of the prosodic structure needed to demarcate the phonological phrases (whose boundaries block assimilation and stress clash avoidance rules) and intonation domains (whose boundaries are marked by deceleration, pause insertion and boundary marking pitch movements). Syntactic structure also determines (in part) which words have to be accented. Finally, lexical category disambiguation is often a by-product of a syntactic parser.

Although the syntactic parser is an important module in any advanced TTS, we take the view that, in principle, its development and evaluation does not belong to the domain of speech output systems. Syntactic parsing is much more a language engineering challenge, needed in automatic translation systems, grammar checking, and the like. For this reason, we refer to the chapters produced by the EAGLES Working Groups on the evaluation of Automatic Translation and Translation tools.

12.5.1.6 Sentence accent

Appropriate accentuation is necessary to direct the listener's attention to the important words in the sentence. Inappropriate accentuation may lead to misunderstandings and delays in processing time (cf. Terken 1985). For this reason most TTS-systems provide for accent placement rules. Accentuation rules can be evaluated at the symbolic and the acoustic level.

⁵On an isolated word basis, this latter category could be considered as an error; however, the error can very often be corrected post-hoc when strings of morphologically segmented words are analysed further by a syntactic parser.

Monaghan and Ladd (1989, 1990) tested the symbolic output of a sentence accent assignment algorithm applied to four English 250 word texts (transcripts of radio broadcasts). The algorithm generated primary and secondary accents, which were rated on a 4-point appropriateness scale by three expert judges. Van Bezooijen and Pols (1989) tested a Dutch accent assignment algorithm at the symbolic as well as the acoustic level (only one type of accent is postulated for Dutch) using 8 isolated sentences and 8 short newspaper texts. Two important points emerged from this study:

- Correlations between the symbolic and the acoustic evaluations were significant but rather low, which means that tests at the symbolic level are not adequate substitute for acoustic tests, and
- Ratings for isolated sentences were more favourable than for sentences in paragraphs, which means that paragraph testing is necessary if the speech output system has to produce connected text.

Again, these are scattered tests, addressing only a handful of the problems that a linguistic module has to take care of. We would recommend the development of a comprehensive test procedure that identifies categories of accent placement error at the sentence and the paragraph level. The principles that underlie sentence accent placement are largely the same across EU languages, so that it makes sense to develop the test procedure on a multilingual basis.

12.5.2 Acoustic aspects

In Section 12.4, speech output assessment was approached within a black box context, i.e. with an emphasis on the speech output as a whole. Black box tests are by force acoustic in nature. In the present section acoustic tests are discussed as well, however this time within a glass box context, which means that attention is focussed on the quality of separate modules, mainly with a view to diagnostic testing. The structure of this section is based on traditional views in phonetics (e.g. Abercrombie 1967) according to which three layers are present in speech: a segmental layer (related to short-term fluctuations in the speech signal), a voice dynamics or prosodic layer (medium-term fluctuations), and a voice characteristics (or voice quality) layer (long-term fluctuations). We will make the same distinction here, 12.5.2.1 being concerned with testing segments, 12.5.2.3 with prosody, and 12.5.2.4 with voice characteristics.

12.5.2.1 Segments

12.5.2.1.1 Functions of segments

The primary function of segments, i.e. the consonants and vowels in the language, is simply to enable listeners to recognise words. Generally, when the segments are sufficiently identifiable, words can be recognised regardless of the durations of the segments and the melodic pattern. In the experience of most researchers good quality (readily identifiable) vowels are afforded by even the simplest speech synthesis systems. One reason is that most coding schemes allow adequate parametrisation of vocalic sounds (narrow band formants slowly varying with time). The synthesis of good quality consonants is an altogether different matter (due to multiple excitation signals, notion of formant not always applicable, abrupt spectral changes), and this is where most (parametric) synthesisers show defects.

Moreover, since speech extends along the time dimension, segments early in the word in practice contribute more to auditory word recognition than later segments. Trailing segments, especially in long (i.e. polysyllabic) words are often not needed to distinguish the word from its competitors. Also, stressed syllables tend to contribute more to a word's identity than segments in unstressed syllables. For these reasons it makes sense to break down the segmental quality of speech output systems for vowels and consonants in various positions (initial, medial, final), within monosyllabic and polysyllabic words, and in stressed versus unstressed syllables.

12.5.2.1.2 Segmental tests

Of all aspects of synthetic speech output, the evaluation of the segmental aspect has received most attention so far, because:

- Good segmental quality is considered to be the main prerequisite for good overall quality.
- There is general agreement about the relevant categories in terms of which quality can be assessed, namely phonemes.
- It is easy to establish accuracy: a phoneme intended to represent /f/ is identified accurately if listeners respond with /f/.

Near perfect segmental quality is essential for applications with a strong emphasis on the transmission of low-predictability information to untrained listeners, for example traffic information services and reverse telephone directory assistance (*What name and address belong to this telephone number?*). Unlike the case of "normal" words, the pronunciation of names cannot be deduced from the context. Moreover, for names it is particularly important that each consonant and vowel be clearly enunciated because there are many near-homophones, i.e. names that differ in just one sound, and strange names which listeners may never have heard before. In applications like these, where prosody is of minor importance, the required intelligibility level can be attained for instance by making use of canned speech or waveform concatenation. In other applications, where text-to-speech is preferred, it may perhaps not be necessary for each sound to be identified correctly. However, since very little is known as yet on the specific contributions of

single sounds to overall intelligibility, synthesis designers have usually taken the pragmatic position that in principle all sounds should be identifiable. In that case detailed diagnostic testing of segmental quality using a glass box approach remains necessary.

As stated above, many tests have been developed to evaluate the quality of synthetic segments. There is a basic distinction between segmental tests at the word level, where single words (meaningful, meaningless or lexically unpredictable) are presented to listeners, and segmental tests at the sentence level, where complete sentences (meaningful, meaningless, or semantically unpredictable) are presented to listeners. Within either category tests can be categorised in functional and judgment studies.

12.5.2.1.2.1 Segmental tests at the word level

FUNCTIONAL SEGMENTAL TESTS AT THE WORD LEVEL

The test approach used to evaluate segments at the word level has been in general functional, quality being expressed in terms of the percentage of correct phoneme identification. In this section we will discuss the Diagnostic Rhyme Test (DRT), the Modified Rhyme Test (MRT), the SAM Standard Segmental Test, an anonymous type of test which we shall henceforth call the Diphone Test, the CLuster IDentification (CLID) Test, the Bellcore Test, and the (Modified) Minimal Pairs Intelligibility Test. The reasons why these tests were selected for inclusion are varied: because they are well-known, well-designed, easy and fast to administer and/or promising. Summary information on most tests is provided in Section 12.7.

DRT (DIAGNOSTIC RHYME TEST) AND MRT (MODIFIED RHYME TEST)

The DRT (see Section 12.7.4) is a closed response test with two response alternatives. Items presented to the subjects are of the form CVC, i.e. an initial Consonant followed by a medial Vowel followed by a final Consonant. The identifiability of the medial vowel and final consonant is not examined, only the identifiability of the initial consonant is tested. All items are meaningful, which means that only factors (1) through (3) as listed in Section 12.4.1.1 can influence the results. In order to obtain insight into the precise nature of possibly poor identifiability of initial consonants, the two categories from among which the subjects are forced to select the correct response contain minimal phonemic contrasts. The subject would be asked for instance to indicate whether a synthetic item was intended as *dune* or *tune*.

The MRT (see Section 12.7.5) is an (originally) closed response test with six response alternatives. All items are of the form CVC and (in its original

form) meaningful. Both the identifiability of initial and of final consonants are tested, but never simultaneously. An example of response alternatives testing the identifiability of a final consonant would be a contrastive syllable coda series such as *peas, peak, peal, peace, peach*, and *peat*. The use of meaningful test items has some positive effects:

- The DRT and MRT are reliable, fast, and easy to administer and score.
- No training is required of subjects because they can respond in normal spelling.

However, the DRT $\,$ and MRT have some serious drawbacks and restrictions as well:

- Intelligibility may be overestimated since subjects adjust their perception to the response categories presented to them.
- There is a risk of ceiling effect.
- Due to their restricted coverage and their limitation to meaningful words, they have little diagnostic value.

Both the DRT and MRT have a long tradition in speech output assessment and have been used in many studies, mainly for comparative purposes. The DRT has been employed among others by Pratt (1987), who compared a wide range of synthetic voices/systems and a human reference, both clear and with noise added to give a speech-to-noise ratio of 0 dB(A). Eight subjects participated. The percentages correct for the human voice and five synthesisers are given in Table 12.4.

Table 12.4: Some results obtained with the DRT by Pratt (1987)

Speech type	Clear	$0 \mathrm{dB}(\mathrm{A})$
Human	96	80
DECtalk Beautiful Betty	92	64
DECtalk Perfect Paul	88	60
Prose 2000 v1.2	81	42
Namal Type and Talk	66	40
Computer Concepts Speech Rom	61	30

All factors (speech system, speech-to-noise ratio, and type of phonemic contrast in the two response categories) had a significant effect on the percentage of correct identification. More interestingly, all interactions appeared to be significant as well. For example, as can be seen above, the intelligibility of synthetic speech was affected by adding noise to a much higher degree than that of human speech. Moreover, adding noise extended the range of the percentages correct, thus making the test more sensitive. So, if synthesis

systems are compared which are rather similar, it might be a advisable to add noise.

The MRT has been employed, among others, by Logan et al. (1985) to evaluate eight synthesisers and a human reference. On the basis of the results, the systems were grouped into four categories, namely (1) human voice, (2) high-quality DECtalk 1.8 Paul, DECtalk 1.8 Betty, MITalk-79, Prose 3.0, (3) moderate-quality INFOVOX SA 101, Berkeley, TSI-proto 1, and (4) low-quality Votrax Type'n'Talk, and Echo. Percentages correct are given in Table 12.5 for the closed response variant.

Table 12.5: Some results for the MRT (Logan et al. 1985)

Speech type	Initial C	Final C	Overall
Human	99	99	99
High-quality	96	93	95
Moderate-quality	90	82	85
Low-quality	66	71	68

The categories distinguished could be used as benchmarks (although the data are somewhat dated, the set of synthesisers tested is probably representative of the quality range of more recent synthesisers). Methodological matters were considered as well. A test/retest design proved the MRT to be reliable. Moreover, the closed and open response variants (compared for the five best systems) yielded the same rank order.

SAM STANDARD SEGMENTAL TEST

For diagnostic purposes the SAM Standard Segmental Test developed within the Speech Assessment Methods (SAM) project of ESPRIT (see Section 12.7.1) is to be preferred above the DRT and MRT tests. The test items in this test consist of meaningless and (sometimes by chance) meaningful, i.e. lexically unpredictable stimuli, which means that factors (1) and (2) as listed in Section 12.4.1.1 have an effect on the responses. Items are CV, VC, and VCV stimuli, where C stands for all consonants allowed in the given position in a given language and V for one of the three point vowels of the given language, typically open /a/, close front /i/, and close back /u/. So, all permissible consonants are tested in word initial, word medial, and word final position. Vowels are not tested; they provide varying phonetic contexts within which the consonants to be tested are placed (the identifiability of sounds can vary depending upon neighbouring sounds). Examples of test items are *pa*, *ap*, *apa*, *ki*, *ik*, and *iki*. An open response format is used, i.e. listeners choose a response from among all consonants.

The SAM Standard Segmental Test has many positive points:

- It is relatively fast.
- It provides precise information on the characteristics of consonants as perceived by listeners, these can be compared to the intended characteristics, thus indicating in which direction improvement should be sought.
- It allows comparison of segmental intelligibility over languages.
- Software is available, allowing automatic stimulus generation and scoring (in fact, the stimulus set is really a subset of the CLID test (see Section 12.7.2), which allows for more vowel contexts and syllable structures).

The main restriction of the SAM Standard Segmental Test is that:

• It is restricted in that only single consonants are tested, while vowels are left out of consideration completely.

Part of the SAM Standard Segmental Test has been applied to English, German, Swedish, and Dutch synthesisers. Comparative results are available for Swedish medial C produced by a human and two synthesisers as perceived by listeners with perfect and imperfect hearing (Goldstein and Till 1992). The percentages correct medial C identification are given in Table 12.6. Of the 54 test items, 3 were found to differ significantly (p=2%) between human and KTH, 9 between human and INFOVOX, and 3 between KTH and INFOVOX.

Table 12.6: Some results for the SAM Standard Segmental Test (Goldstein and Till 1992)

Speech type	Perfect hearing (N=24)	Imperfect hearing $(N=14)$
Human	94	91
m KTH-synthesis	91	84
INFOVOX-synthesis	88	79

DIPHONE TEST

A more complete overview of the performance of segments in a wider variety of contexts is provided by a test which assesses the intelligibility of all permissible (pronounceable) CVC, CVVC, VCV, and VCCV sequences of a given language. Such a test will be referred to as a Diphone Test, because the test items can be constructed by combining all the diphones in a diphone inventory. Just like in the SAM Standard Segmental Test, the test items are lexically unpredictable and the response categories open, so that it is useful for diagnostic purposes. Extra advantages of the Diphone Test over the SAM Standard Segmental Test are the following:

• The Diphone Test is less restricted than the SAM Standard Segmental Test because in addition to consonants vowels are tested,

- 524 Part III: Spoken language system assessment
 - Both monosyllabic (CVC) and disyllabic (CV-VC, V-CV, VC-CV) structures are represented,
 - Due to the VCCV items, heterosyllabic consonant clusters are included, i.e. sequences of consonants belonging to different syllables.

The main disadvantages/restrictions of the Diphone Test are:

- It is not incorporated in a software package,
- It does not comprise tautosyllabic consonants, i.e. sequences of consonants within the same syllable.

The Diphone Test has been used to evaluate diphone synthesis in French (Pols et al. 1987), Italian (Van Son et al. 1988), and Dutch (Van Bezooijen and Pols 1987). The Dutch Diphone Test combined all Dutch diphones into a set of 768 test items: 307 CVC, 173 VCV, 267 VCCV, and 21 CVVC. The only thing needed to construct the test material for a particular language is a matrix with the phonotactic constraints operating in that language, i.e. restrictions on the occurrence of all consonants and vowels in various word positions and phonetic contexts. Such matrices have been constructed for a number of European languages within the ESPRIT-SAM project.

BELLCORE TEST AND CLID (CLUSTER IDENTIFICATION) TEST

As mentioned above, even the Diphone Test is not complete, since no tautosyllabic consonant clusters are included. The importance of this structure should not be underestimated. According to Spiegel et al. (1990), about 40% of all one-syllable words in English begin and 60% end with consonant clusters. The Bellcore Test (see Section 12.7.3) and the CLID Test (see Section 12.7.2), have been developed to fill this gap.

The Bellcore Test has a fixed set of CVC-stimuli, comprising both meaningless and meaningful words, e.g. *frimp* and *friend* or *glurch* and *parch*. Tautosyllabic consonant clusters and single clusters are tested separately in initial and final position. Vowels are not tested. Open response categories are used. Compared to the Diphone Test, the Bellcore Test has some advantages, the main being that:

• It includes tautosyllabic consonant clusters.

However, the Bellcore Test has restrictions as well:

- It is confined to monosyllabic structures.
- Vowels are not tested.
- No test material is available for other languages than English.

The test has been applied to assess the intelligibility of two synthesisers compared with human speech, presented over the telephone (Spiegel et al.

1990). The syllable score was 88% for human telephone speech and around 70% for the synthetic telephone speech. Consonant clusters had lower intelligibility than single consonants. Intelligibility for meaningful words was higher than for meaningless words, a finding which could not be explained. The CLID Test is a very flexible architecture which can be used for generating a wide variety of monosyllabic test items in an in principle unlimited number of languages. Both meaningful and meaningless items can be generated as long as matrices are available with the phonotactic constraints to be taken into account. Open response categories are used. Intelligibility can be assessed in whatever way one wants. The CLID test has been applied to testing the intelligibility of German synthesisers (e.g. Jekosch 1992; Kraft and Portele 1995).

The CLID test had all the advantages of the SAM Segmental Test listed above, whereas it does not have the restriction mentioned. Thus the positive points of the CLID test are the following:

- It is relatively fast.
- Clusters of consonants (tautosyllabic and heterosyllabic) and single consonants are tested as well as clusters of vowels and single vowels.
- It provides detailed diagnostic information.
- It allows comparison of segmental intelligibility over languages.
- Software is available, allowing automatic stimulus generation and scoring.

(Modified) Minimal Pairs Intelligibility Test

The last tests we want to mention in this context are the so-called Minimal Pairs Intelligibility Test (MPI Test), proposed by Van Santen (1993) as an alternative to the DRT, and a modification to it introduced by Syrdal and Sciacca (1994), the Diagnostic Pairs Sentence Intelligibility Evaluation Test (DPSIE Test). These tests were designed to reduce ceiling effects and expand the coverage of the DRT to include:

- vowels
- tautosyllabic (same syllable) as well as heterosyllabic (different syllable) consonant clusters,
- unstressed syllables,
- · de-accented or cliticised words,
- words in sentences,
- polysyllabic words,
- insertions and deletions.

The MPI Test consists of a fixed set of 256 sentence pairs containing one contrast, e.g. *The horrid courts score a revolution* versus *The horrid courts score a revolution*. The minimal pair appears on the screen and the correct sentence has to be identified. Differences between the MPI Test and the DPSIE Test include:

- A change of binary Jakobson et al. (1951) distinctive features to multivalued Ladefoged (1975) distinctive features, so that more realistic perceptual confusions can be tested.
- DPSIE generates its word pairs matched for familiarity (i.e. token frequency) and verb transitivity.

The main advantage of the MPI and DPSIE Tests is that:

• The intelligibility of segments is evaluated in a virtually complete variety of conditions and phonetic contexts, potentially providing a host of diagnostic information.

The main disadvantages of the tests are:

- They are inefficient, since each response gives information on the identifiability of only one phoneme.
- Inclusion of segments in stimuli is restricted by their occurrence in the language.
- Creating test materials presupposes the availability of large databases.

J udgment tests at the word level

In principle, in addition to functional intelligibility tests, judgment tests, where subjects rate their subjective impression of the stimuli on scales, are possible for evaluating the segmental quality at the word level as well. For example, Van Bezooijen (1988), in addition to running a consonant cluster identification test, presented 26 Dutch consonant clusters (both initial and final) to be rated on naturalness, intelligibility, and pleasantness. The clusters were embedded in meaningful words. In order to obtain "pure" judgments, unaffected by the quality of the rest of the word, subjects were explicitly asked to pay attention to the clusters only. So, the test required analytic listening. However, one can never be sure to what extent listeners in fact stick to the instructions. Perhaps this is one of the reasons why judgment tests of this type have been rare.

12.5.2.1.2.2 Segmental tests at the sentence level

In addition to the word level, tests for the assessment of segmental quality have been developed at the sentence level as well. Here the effect of prosody could be minimised by presenting the material on a monotone, but in practice, if only for naturalness' sake, prosody is usually included. Compared with segmental tests at the word level, tests at the sentence level are more similar to speech perception in normal communication but at the same time, as a consequence, less suitable for diagnostic purposes, for the following reasons:

- With sentences, the intelligibility scores will not only be based on segmental quality but also to some extent on prosodic quality, so that poor intelligibility is more difficult to trace back to specific sources.
- The composition of the test material is somewhat unsystematic, so that no complete confusion matrices can be obtained.
- Especially with semantically normal sentences listeners will not only rely on segmental information but use other information sources as well, related to word combinatory probabilities, and, in the case of meaningful sentences, to semantic coherence (factors (4) and (5) in Section 12.4.1.1).

Of course, if the test is not intended as a diagnostic tool but has a purely comparative aim, these consequences of using sentences do not necessarily detract from its value. However, it is important to remember that as soon as complete sentences are presented to listeners, the test is no longer limited to evaluating segmental quality alone. This means that the title of this section "segmental tests at the sentence level" is not completely adequate. In fact, depending on the extent to which restrictions are imposed on the construction of the test materials, tests at the sentence level are in between a glass box approach and a black box approach. So, the main differences among the segmental sentence tests described is their position on the glass box – black box continuum.

In this section only functional tests will be discussed. In addition, judgment tests at the sentence level have frequently been carried out. These are described under the heading "black box approach" in Section 12.4.1.3, where judgment tests to evaluate *overall output quality* are discussed. These tests entail the rating of scales such as *acceptability, intelligibility* and *naturalness*.

HARVARD PSYCHOACOUSTIC SENTENCES

One of the most well-known segmental intelligibility tests at the sentence level is the fixed set of 100 semantically and syntactically "normal" Harvard Psychoacoustic Sentences (Add salt before you fry the egg) (see Section 12.7.8). Intelligibility is expressed by means of the percentage of correctly identified keywords (nouns and verbs). In this test no restrictions are placed upon the composition of the test materials, which means that the percentage of correct responses is determined only to a limited extent by the acoustic characteristics of the individual segments. This test would therefore have to be placed towards the black box end of the continuum. In terms of the factors listed in Section 12.4.1.1, only (6), (7) and (8) are excluded.

The main advantages of the Harvard Psychoacoustic Sentences Test are:

- It is easy to administer (no training required of the subjects).
- Responses are easy to score (even manually).

The main disadvantage/restrictions of the test are:

- Limitation to comparative purposes.
- A strong learning effect (the same sentences can only be used once).
- Restricted number of test items.
- Restricted generalisability, because there is only one syntactic structure.
- Danger of a ceiling effect (so only fit to be used for low-quality systems).

The Harvard Psychoacoustic Sentences were compared with the Haskins sentences by Pisoni et al. (1985a,b) for four synthesisers and human speech (see below).

HASKINS SYNTACTIC SENTENCES

Another famous test at the sentence level is the fixed set of 100 Haskins Syntactic Sentences (see Section 12.7.6). These sentences are semantically unpredictable, which means that they do not occur in daily life. An example is *The old farm cost the blood*. In terms of advantages and disadvantages, the Harvard Sentences and Haskins Sentences have much in common. The only difference is that Haskins listeners can rely less on semantic coherence (factor (5) in the list of factors in Section 12.4.1.1) so that the role of the acoustic characteristics of the segments is more important. Therefore, the Haskins sentences find themselves somewhat closer to the glass box end of the continuum than the Harvard sentences. The Haskins sentences were applied to four synthesisers and human speech by Pisoni et al. (1985a,b), and were compared with the Harvard sentences. The percentage of correct keyword identification is given in Table 12.7.

 Table 12.7: Some results for the Harvard Psychoacoustic Sentences

 and Haskins Syntactic Sentences

Speech type	Harvard (meaningful)	Haskins (unpredictable)
Human	99	98
DEC Paul v1.8	95	87
MITalk-79	93	79
DEC Betty v1.8	90	75
Prose-2000 prototype	84	64

It can be seen that the two tests yield the same rankorder. However, as expected, due to the reduced semantic coherence, the Haskins sentences are more sensitive.

SEMANTICALLY UNPREDICTABLE SENTENCES (SUS)

Both the Harvard and Haskins sentences had a fixed set of sentences, characterised by a single syntactic structure, which could be used as test materials. More recently, a more flexible approach was opted for with Semantically Unpredictable Sentences (SUS), developed by SAM (see Section 12.7.7). The test materials in the SUS consist of a fixed set of five syntactic structures which are common in most Western European languages, such as "Subject– Verb–Adverbial" (*The table walked through the blue truth*). The lexical slots in these structures are filled with high-frequency words from language specific lexicons. The resulting stimulus sentences are semantically unpredictable, just like the Haskins Syntactic Sentences.

The advantages of the SUS Test are the following:

- Easy to administer (no training required of the subjects).
- Responses can be scored automatically.
- Reasonable generalisability due to the variation in syntactic structures and lexical items.
- Comparability across languages.

The main disadvantages/restrictions of the test are:

- Limited to comparative purposes.
- Learning effects (despite the large number of test items).
- Risk of ceiling effects (so only fit to be used for low-quality systems).

Pilot studies with the SUS test have been run in French, German, and English (Benoît 1989; Benoît et al. 1989; Hazan and Grice 1989). Results showed, among other things, that keywords presented in isolation were identified significantly less well than the same words in a sentence context. This is attributed in part to the fact that the syntactic category of the isolated words is not known. Furthermore, the SUS were found to be sensitive enough to discriminate between two synthesisers differing in prosody.

12.5.2.2 Recommendations on specific segmental quality tests

- 1. Use the CLID Test for the evaluation of the segmental intelligibility at the word level, both for diagnostic and comparative purposes (in the latter case the stimulus set can be smaller).
- 2. Use the SUS Test to evaluate intelligibility for comparative purposes at the sentence level.

12.5.2.3 Prosody

12.5.2.3.1 Functions of prosody

By prosody we mean the ensemble of properties of speech utterances that cannot be derived in a straightforward fashion from the identity of the vowel and consonant phonemes that are strung together in the linguistic representation underlying the speech utterance. Prosody would then comprise the

melody of the speech, word and phrase boundaries, (word) stress, (sentence) accent, rhythm, tempo, and changes in speaking rate. We exclude from the realm of prosody the class of voice characteristics (see Section 12.5.2.4). Prosodic features may be used to differentiate between otherwise identical words in a language (e.g. trusty - trustee, or export (noun) - export (verb), with initial stress versus final stress, respectively). Yet, word stress is not so much concerned with making lexical distinctions (this is what vowels and consonants are for) as with providing checks and bounds to the word recognition process. Hearing a stressed syllable in languages with more or less fixed stress informs the listener where a new word may begin; error responses in word recognition strongly tend to agree with the stimulus in terms of stress position. In a minority of the EU languages (Swedish, Norwegian) lexical tone (rather than stress) is exploited for the purpose of differentiating between segmentally identical words (See also Chapter 6). The more important functions of prosody, however, are located at the linguistic levels above the word:

- Prosody offers explicit segmentation cues in the form of phrase boundaries, i.e. it tells the listener which words go together and should be interpreted as making up a coherent chunk of information; also, these cues allow the listener to determine the "depth" of the break between chunks, i.e. whether the end of a word group, clause, sentence, or even a whole paragraph has been reached.
- Prosody provides an indication for the listener which words are presented by the speaker as expressing important information (highlighting or focussing through accentuation).
- Prosody, especially melody, carries some meaning of its own (intonational meaning) which, for example, allows the speaker to present a sentence as a statement or a question, or to express emotions and/or attitudes towards the verbal contents of the message or towards the hearer.

These functions suggest that proceedy affects comprehension (establishing the semantic relationships between words) rather than intelligibility (determining the identity of words), and, indeed, this is what most functional tests of proceedy aim to evaluate.

12.5.2.3.2 Judgment tests of prosody

Evaluation of the prosody of speech output systems is alternately focussed on the formal or the functional aspects. Only a handful of tests are directed at the formal quality of temporal organisation. An exemplary evaluation study on the duration rules of MITalk (Allen et al. 1987) was done by Carlson et al. (1979). They generated six different versions of a set of sentences by including or excluding effects of consonant duration rules, vowel duration rules, and stressed syllable and preboundary lengthening rules in the synthesis. These versions were compared with a topline reference condition where (normalised) segment durations copied from human versions of the test sentences spoken by the MITalk designated talker were imposed on the synthesis. There were two baseline conditions, one with the neutral (inherent) table values substituted for all segments, and one with random segment duration variation (within realistic bounds). The results showed that the temporal organisation afforded by the complete rule set was judged to be natural as the human topline control. Moreover, sentences generated with boundary markers at minor and major breaks were judged to be more natural than speech without boundary markers.⁶

More work has been done in the field of melodic structure. Let us first consider judgments of formal aspects of speech melody. The formal properties of, for example, pitch movements or complete speech melodies can be tested by asking groups of listeners (either naive or expert) to state their preference in pairwise comparisons or to rate a melody in a more absolute way along some goodness or naturalness scale. At the level of elementary pitch movements (such as accent-lending or boundary marking rises, falls, or rise–fall combinations) the SAM Prosodic Form Test (see Section 12.7.9) is a useful tool. The test was applied to two English and two Italian synthesisers, with 3 contours, 4 levels of segmental complexity, 5 items at each level, 4 repetitions of each token (Grice et al. 1991). Significant effects were found for synthesiser and contour, as well as for the interactions between synthesiser and contour and between synthesiser, complexity, and contour. By relating the scores for the contours to those for the monotone reference, the effect of differences in segmental quality on the ratings could be cancelled out.

Using the same methodology, i.e. rating and pairwise comparisons, the quality of synthetic speech melody can be evaluated at the higher linguistic levels. At the level of isolated sentences pairwise comparisons of competing intonation-by-rule modules is feasible when the number of systems (or versions) is limited (e.g. Akers and Lennig 1985). When multiple modules are tested using a larger variety of sentences and melodies, scale rating is to be preferred over pairwise comparisons for reasons of efficiency (De Pijper 1983; Willems et al. 1988).

Evaluation of speech melody generators should not stop at the level of isolated sentences. Ratings by expert listeners in Dutch could not reveal any quality differences between synthetic melodies and a human reference when

⁶Later (cf. Allen et al. 1987), the duration rules were evaluated directly (objectively) by comparing the predicted segment durations with the segment durations as measured in spectrograms of new paragraphs read by the designated speaker. The rules accounted for 84% of the duration variance with a residual standard deviation of 17 ms (excluding the prediction of pause duration). Seventeen ms is generally less than the just noticeable difference for a duration change in a single segment in a sentence context (Klatt 1976), which would explain why the human reference and the rule-derived durations were judged equally natural.

the sentences were listened to in isolation; however, the same synthetic melodies proved inferior to the human reference when they were presented in the context of a full paragraph (Terken and Collier 1989). Along the same lines, Salza et al. (1993) evaluated the prosody of the Eloquens TTS for Italian in a train schedule consultation application. Three hundred sentences were tested, realistically distributed over seven melodic modalities: Command sentences, Simple declaratives, List sentences, Wh-questions, Yes/no-questions, Yes/no-echo questions, and Yes-no modal questions. Expert listeners' scores did not differ from those of naive subjects, and scores were better for utterances presented as part of a dialogue than for sentences presented in isolation. Clearly, both studies demonstrate that paragraph position or function within a dialogue induces certain perceptually and communicatively relevant adaptations to sentence prosody.

The form tests discussed so far address prosody globally. An analytic approach to prosodic evaluation using judgment testing was proposed by Bladon (1990) and co-workers. They developed an elaborate check list of formal properties that should be satisfied by any speech output system that claims to generate English melodies. Trained (but phonetically naive) judges listen to synthetic utterances, while looking at orthographic transcripts of the utterance with a crucial word or syllable underlined. Their task is to check whether the target syllable does in fact contain the melodic property prescribed by the check list. Although this idea is attractive from a diagnostic point of view (melodic flaws are immediately identified) the system has some drawbacks that should be considered before extending its use to other materials and other languages. First, drawing up a valid checklist presupposes a theory of intonation, or at least a detailed and valid description of the test sentences. Workable theories and descriptions may be provided for English and some other languages, but will not be available for all (EU) languages. Second, even for English, the criteria for each melodic check point were formulated in rather crude terms, which makes it difficult for the judges to determine whether the utterance does or does not satisfy the criterion. Third, it is impossible to determine the overall quality of the melodies tested, since there is no way of combining the pass/fail scores for the various check points into a weighted overall score. A preliminary experiment revealed that three output systems could be meaningfully rankordered along a quality scale, but not at the interval measurement level. Systems that were clearly different as judged by experts were very close to each other in terms of their unweighed overall score, whereas systems that were rated as equally good by experts differed by many points.

For the reasons given above, we do *not* recommend analytic judgments by naive listeners using a check list as an approach to evaluating prosody.

There is (at least) one judgment test that assesses how well certain communicative functions are signalled by prosody at a higher level. The SAM Prosodic Function Test⁷ (see Section 12.7.9) asks for ratings of the communicative appropriateness of melodies in the context of plausible humanmachine dialogue situations. The test was applied to human-machine dialogues designed to simulate a telephone enquiry service giving flight information (Grice et al. 1992b). A restricted set of contexts and illocutionary acts were included: asking (seeking information, seeking confirmation), assertive (conclude, put forward, state), expressive (greet), and commissive (offer, propose to). Two intonation versions were compared, one based on an orthographic input with punctuation (target intonation algorithm) and the other based on a text input edited to conform to the type of text generated by an automatic language generator (reference intonation algorithm). The test should be seen as a first attempt to evaluate the paralinguistic appropriateness of intonation in dialogue situations. For general comparative purposes, it would be useful to have an agreed-on, systematic inventory of situations or speech acts one would want to include, taking as a point of departure, for example, the classification of speech acts proposed by Searle (1979).

Finally, we are not aware of tests asking subjects to judge the quality of the expression of emotions and attitudes in synthetic speech. It would appear that functional testing of these qualities is preferred in all cases.

12.5.2.3.3 Functional tests of prosody

Evaluating speech output prosody using functional tests is even more in its infancy. Since prosody is highly redundant given the segmental information (with the exception of the signalling of sentence type and emotion/attitude), it can be functionally tested only if measures are taken to reduce its redundancy. The first course of action, then, has been to concentrate on atypical, rather contrived materials in which prosody is non-redundant. That is, the materials consist of segmental structures that would be ambiguous without the prosody, and listeners are asked to solve the ambiguity. To the extent that the disambiguation is successful, the speech output system can be said to possess the appropriate prosodic functions. We find examples of such functional tests for the disambiguation of minimal stress pairs (for a survey, see Beckman 1986), word boundaries (for a survey, see Quené 1993), constituent structure (syntactic/prosodic bracketing, e.g. Lehiste et al. 1976), sentence type (e.g. Thorsen 1980), and focus distribution (e.g. Nooteboom and Kruijt 1987). However, in these kinds of study speech output assessment typically was not the primary research goal. Rather, speech synthesis

 $^{^{7}}$ The notion "function test" in this sense has no relationship with our use of the term "functional test". In the SAM Prosodic Function Test prosodic quality is not being tested in a functional task: we are still dealing with intuitive judgments (ratings) of how well the melody would fulfil its function without actually testing it.

was used here by psycholinguists or experimental phoneticians to manipulate the speech parameters in a controlled fashion.

The second route is to make prosody less redundant by degrading the segmental quality, such that without prosody (i.e. in the baseline conditions identified above) the intelligibility of the speech output would be extremely poor. The quality of the prosody would then be measured in terms of the gain in intelligibility, i.e. increase in the percentage of correctly reported linguistic units (phonemes, morphemes, words) due to the addition of prosody. Carlson et al. (1979) measured intelligibility of utterances synthesised by MITalk with and without application of vowel duration, consonant duration and boundary marking rules (see above). They found that adding duration rules improved word intelligibility; adding within-sentence boundaries, however, did not boost intelligibility (even though the result was judged to be more natural, see above). Scharpff and Van Heuven (1988) demonstrate that adding within-sentence boundaries (i.e. changing the temporal organisation) does improve word intelligibility (especially for monosyllabic words) in Dutch diphone synthesis, and that utterances with pauses were judged as more pleasant to listen to (but only when listeners were unfamiliar with the contents of the sentence (Van Heuven and Scharpff 1991). Reasoning along the same lines, one would predict that quality differences in speech melody would have an effect on word recognition in segmentally degraded speech. Such effects were, in fact, reported by Maassen and Povel (1985), who used (highly abnormal) speech utterances produced by deaf speakers resynthesised with corrected temporal and/or melodic organisation.

There is a substantial literature on the perception of emotion and attitude in human speech (for a survey, see Murray and Arnott 1993). Typically, listeners are asked to indicate which emotion they perceive in the stimulus utterance, in open or closed response format. Predictably, the larger the set of response alternatives, the poorer the identification of each emotion. It is not clear, in this context, how many different emotions should be distinguished, and to what extent these can be signalled by phonetic means. Still, results tend to show that the most basic emotions can be identified, in lexically neutral utterances, at better than 50% correct, in a 10 alternative closed response test. Synthesis of emotion in speech output is being attempted by several research groups. Preliminary evaluation of emotionby-rule in Dutch diphone synthesis was presented by Vroomen et al. (1993), as summarised in Table 12.8:

- 12.5.2.4 Voice characteristics
- 12.5.2.4.1 Functions of voice characteristics

Whereas the segmental and prosodic features of speech are continuously varying, voice characteristics are taken to refer to aspects of speech which

Table 12.8: Percentage of correctly recognised emotions-by-rule in Dutch diphone synthesis (two diphone sets, obtained from different speakers) and in human speech (after Vroomen et al. 1993)

	neutral	joy	boredom	anger	$\operatorname{sadness}$	fear	indignation
set 1	73	73	90	42	40	35	69
set 2	92	50	98	60	54	46	67
human	66	50	97	66	84	78	88

Note: The human reference condition were neutral utterances with temporal and melodic organisation copied from emotional utterances of the same sentences, using PSOLA.

generally remain relatively constant over longer stretches of speech. Voice characteristics, also referred to as voice quality (cf. Laver 1991), can most easily be viewed as the background against which segmental and prosodic variation is produced and perceived. In our definition, it includes such varied aspects of speech as mean pitch level, mean loudness, mean tempo, harshness, creak, whisper, tongue body orientation, dialect, accent, etc. Voice quality is mainly used by the listener to form a (sometimes incorrect) idea of the speaker's

- mood and personality (cheerful, reliable, dominant),
- physical size (tall, large, strong),
- sex (male, female),
- age (child, young adult, aged),
- regional background (globally "from the North" or more precisely "from London, Paris, or New York"),
- socio-economic status (high/low education),
- *health* (e.g. cold),
- *identity*.

In principle voice quality is not communicative, i.e. not consciously used by the speaker to make the listener aware of something which he was not previously aware of, but informative, which means that, regardless of the intention of the speaker, it is used by the listener to infer information (see also Chapter 11). This information may have practical consequences for the continuation of the communicative interaction, since it may influence the listener's attitudes towards the speaker in a positive or negative sense and may affect his interpretation of the message (cf. Laver 1994).

Recently, increased attention has been paid to voice quality aspects of synthetic speech. In fact, Sorin (1994) regards the successful creation of personalised synthetic voices ("personalised TTS") as one of the most ambitious

challenges of the near future. This aspect of synthesis is, for example, relevant in such applications as Translating (Interpreting) Telephony services, where along with translating the content of the message the original voice of the speaker has to be reconstructed (automatic voice conversion). Moreover, the correct encoding of speaker characteristics such as sex, age, and regional background is also relevant for the synthetic reading of novels for the blind. Finally, a third application is to be found in non-speaking disabled individuals, who have to use a synthetic speech to replace their own. With a view of the latter application, Murray and Arnott (1993) describe a system allowing rapid development of new voice "personalities" for the DECtalk synthesiser with immediate feedback to the user. Voice alteration is done by interpolating between the existing DECtalk voices (five male voices, five females voices, and a unisex child). Thus a voice may be created that sounds "a bit like Paul with a bit of Harry". A somewhat different approach aimed at a somewhat different type of application is described by Yarrington and Foulds (1993), who use original recordings of speakers who know they are going to lose their voice to construct speaker-specific diphone sets.

12.5.2.4.2 Voice characteristics tests

Apart from specific requirements imposed by concrete applications, a general requirement of the voice quality of synthetic output is that it should not sound unacceptably unpleasant. Voice pleasantness is one of the scales included in the overall quality test proposed by the ITU-T to evaluate synthetic speech transmitted over the telephone (see Section 12.7.12). It has also been used by Van Bezooijen and Jongenburger (1993) in a field test to evaluate the functioning of an electronic newspaper for the blind. In this test, 24 visually handicapped rated the pleasantness of voice of two synthesisers on a 10-point scale (1: extremely bad, 10: extremely good). Ratings were collected at three points in time: (1) in a first confrontation with the synthesis output, (2) after one month, and (3) after two months of "reading" the newspaper. Interestingly, the pleasantness of voice ratings were found not to change over time, in contrast to the intelligibility ratings, which reflected a strong learning effect. From this it was concluded that voice quality has to be good right from the start; one cannot count on the beneficial effect of habituation. Both synthesis systems were generally considered good enough for the reading of popular scientific books and newspapers. However, partly due to the unpleasant voice quality, they were found unfit for the reading of novels or poetry (Jongenburger and Van Bezooijen 1992). So, voice quality mainly seems to play a role when attention is directed to the form of the message, for recreational purposes. Finally, we hypothesise that perhaps more than for aspects of speech affecting comprehension, motivation and positive attitude might compensate for poor voice quality.

Of course, judgment studies such as these can only provide global information; if results are negative, no diagnostic information is available as to what voice quality component should be improved. There are no standard tests to diagnostically evaluate the voice quality characteristics of speech output. This type of information could in principle be obtained by means of a modular test, where various acoustic parameters affecting voice quality are systematically varied so that their effect on the evaluation of voice quality can be assessed. This would be the most direct approach.

A more indirect approach would involve asking subjects to listen analytically to and rate various aspects of voice quality on separate scales. A potentially useful instrument for obtaining a very detailed description is the Vocal Profile Analysis Protocol developed by Laver (1991). This protocol, which comprises more than 30 voice quality features, requires extensive training. If data are available for several synthesis outputs the descriptive voice quality ratings could be used to predict the overall pleasantness of voice ratings.

It may also be possible to use untrained listeners, although the number of aspects described will necessarily be more limited and less "phonetic". Experience with human speech samples representing various voice quality settings (Van Bezooijen 1986) has shown that naive subjects can reliably describe 1minute speech samples with respect to the following 14 voice quality scales: warm-sharp, smooth-rough, low-high, soft-loud, nasal-free of nasality, clear-dull, trembling-free of trembles, hoarse-free of hoarseness, full-thin, precise-slurred, fast-slow, accentuated-unaccentuated, expressive-flat, and fluent-halting. Again, if descriptive ratings of this type were available for synthetic speech they could be correlated with global ratings of synthesised voice quality. Alternatively, this type of scale could also be used more directly for diagnostic purposes, i.e. subjects could be asked to rate each of these voice quality aspects on a 10-point scale, with 1: extremely bad and 10: extremely good.

However, as mentioned above, experience with detailed perceptual descriptions of voice quality is as yet limited to non-distorted human speech. It remains to be assessed whether such descriptions can also be reliably made for synthetic speech. And even if this proved to be the case, the translation of the results obtained to actual system improvement is not unproblematic, since not much is known about the acoustic basis of perceptual ratings. Attempts in this direction have been rather disappointing (e.g. Boves 1984). In addition to judgment tests to evaluate the formal aspects of voice quality, functional tests may be used to assess the adequacy of voice quality. Al-

though here also no standard tests are available, the procedures are rather straightforward and dictated directly by application requirements. One can

think, for example, of tests in which subjects are asked, in an open or closed response format, to identify the speaker. This would be useful in an application where one tries to construct a synthetic voice for a given speaker or reconstruct the natural voice of a given speaker. Or one can ask people to identify the speaker's sex, or estimate his age or other characteristics.

In this context, accent and dialect features are relevant as well. For example, for Dutch a new set of diphones was derived from a western speaker, because some non-speaking users complained that the old diphone set had too much of a southern accent to be acceptable for communication in their living environment. To test whether naive listeners were in fact able to discriminate between the two diphone sets, listeners from different parts of the Netherlands rated CVC, VCV, and VCCV stimuli produced with the two systems on a 10-point bipolar *regional accent – standard Dutch* scale. The diphone sets were indeed clearly discriminable (Van Bezooijen 1988).

Summarising it can be stated that very little experience has as yet been gained with the diagnostic and comparative evaluation of voice quality of speech output systems, either by means of judgment or functional tests. Moreover, except for specific applications where synthesis is closely connected with the identity of a speaker (in a clinical or automatic voice conversion setting), it is not even clear how much importance is attached to voice quality by naive listeners. How much does it really bother people when voice quality is unpleasant? For example, does an unpleasant voice quality prevent them from using a synthetic information service? It is too early to give concrete recommendations on how to approach the evaluation of voice quality aspects of speech output; this is one of the topics for further investigation in the near future.

12.5.2.5 Relationships among tests

Knowledge about the relationships among tests is important for at least two reasons:

- It allows a better interpretation of the meaning and validity of the test results obtained.
- It can be used to decide upon the test suite which gives a complete picture of all relevant aspects of speech output without being redundant; it is no use employing two tests which (to a large extent) yield the same information.

What would be needed to assess the relationships among tests is a large scale study which compares the performance of all "serious" tests testing the same aspect (e.g. intelligibility or comprehension) for a wide range of synthesisers. One would then like to know the stable differences among the tests in quality measured (e.g. percentage correct), as well as the correlations among the rank orderings of the synthesisers. In addition, it would be useful to have information on the reliability of "identical" tests developed and applied to a wide variety of different languages.

Some differences between the results obtained with different tests can be predicted to some extent. For example, when considering intelligibility, we think at least four factors will affect the outcomes: Intelligibility can be expected to increase

- as the unit of measurement is smaller (it is easier to identify one phoneme correctly than a sequence of phonemes),
- as the structure of the test items is more predictable (fixed versus open structure),
- as the combination of phonemes is more predictable (meaningful versus meaningless),
- as the number of response categories is smaller (closed versus open).

These predictions can be tested by looking at actual intelligibility results. Jekosch and Pols (1994), for example, assessed the intelligibility of one German synthesiser by means of four different tests (all described in Section 12.7):

- The SAM Standard Segmental Test (open response).
- A German variant of the MRT (closed response).
- The CLID test (open response).
- The SUS test (open response).

The percentage of correct elements (phonemes in the SAM Standard Segmental Test, clusters in the CLID test, words in the MRT and the SUS test) differed widely, from 19% to 85%. The lowest percentage was obtained for the SUS test, followed by the SAM Standard Segmental Test, the CLID test, and the MRT. The fact that the highest score was obtained with the MRT agrees with our predictions, since this test possesses not a single aspect with a negative effect on intelligibility: The unit of measurement is small (phoneme), the structure is fixed (CVC), the items are meaningful, and the response set is closed (six categories). The results for the other three tests point to complex interactions among the four factors.

Delogu et al. (1992a) compared four different test methods for evaluating the overall quality of Semantically Unpredictable Sentences produced by a male speaker (once with and once without noise added), three synthesisers, and three vocoders:

- Categorical estimation (5-point scale from "excellent" to "bad").
- Magnitude estimation (quality expressed by any number between 1 and 100).
- Paired comparison ("which of the two realisations do you prefer?").
- Word monitoring ("push a button as soon as a specified target word is recognised").

Very high correlations were obtained among categorical estimation, magnitude estimation, and paired comparison (r > 0.90); somewhat lower but still high correlations were found between these three test methods and reaction time (r around 0.80). Reaction time showed the smallest variation in the responses, but the least discriminatory power. The best discrimination among the systems was obtained with paired comparisons.

Silverman et al. (1990) compared the results of the Bellcore intelligibility test (see Section 12.7.3) with a comprehension test in which subjects had to answer questions related to the content of synthesised utterances with yes, no, or can't tell from the information provided. The faster subjects answered questions, the more items they heard. Two synthesisers, A and B, were tested. The intelligibility test yielded higher percentages correct for A than for B (77% versus 70%), whereas the comprehension test yielded higher percentages correct for B than for A (69% versus 63%). A few remarks are in order when attempting to interpret these seemingly contradictory results:

- 1. The presence of prosody in the comprehension test may have played a role.
- 2. Silverman et al. rightly state that the intelligibility scores relate to utterance-initial and utterance-final phonemes only, which hardly occur in sentence-length material (and running speech in general, for that matter).
- 3. It appeared that subjects generally heard fewer questions from B than from A in the test time allotted, a finding the meaning of which is complicated by the fact that B spoke at a slower rate than A.

Whatever the exact basis of the opposite rank orders yielded by the two tests, it is clear that caution should be exercised when generalising from a laboratory-type intelligibility test to a field-type, application-oriented comprehension test. Low correlations between intelligibility (MRT) and comprehension are also reported by Ralston et al. (1991).

In general, studies comparing different tests comprehend only a limited number of systems, which makes it difficult to learn to what extent the different tests rank the systems in the same way. Moreover, the relationship between the results yielded by glass-box and black-box tests deserves more systematic attention. We think that the importance of further studies of the relationships among tests cannot be stressed enough, if one wants to have a good idea of the meaning and generality of results obtained.

12.6 Further developments in speech output testing

12.6.1 Introduction

In this final section we will consider desired developments in the field of speech output testing, and discuss possibilities for further research on a more general level than we did in the preceding sections. The section consists of three parts. In the first part (Section 12.6.2) we will be concerned

with possibilities of producing more efficient output testing techniques. The general, longer-term strategy proposed here is to replace expensive, timeintensive tests (involving human listeners in field situations) by cheaper, automated tests carried out in a laboratory setting. In order to make this feasible we will have to establish predictability relationships between the various types of tests discussed in Section 12.2:

- How to predict human responses to output systems from automated (computer-simulated) listeners? (Section 12.6.2.1)
- How to predict overall performance assessment from evaluation results of system modules? (Section 12.6.2.2)
- How to predict functional performance of a system from judgment testing? (Section 12.6.2.3)
- How to predict field performance of an output system from laboratory testing? (Section 12.6.2.4)

Next, we will discuss (Section 12.6.3) developments that we feel are needed in the assessment of linguistic interfaces of speech output systems. Finally, in Section 12.6.4, we will propose research for the mid term aimed at improving speech output evaluation at the acoustic level in each of the four areas identified: segmental quality (Section 12.6.4.1), prosodic quality (Section 12.6.4.2), voice characteristics (Section 12.6.4.3), and overall quality (Section 12.6.4.4).

12.6.2 Long-term strategy: Towards predictive tests

12.6.2.1 From human to automated testing

The ultimate criterion to decide on the quality of speech output resides with the human listener. Speech output assessment is therefore basically a matter of human perception research. It is commonly acknowledged that the human listener is a noisy measurement instrument, which causes output assessment to be a slow and (therefore) expensive undertaking. There are generally felt to be two ways out of this problem. One is to look for assessment procedures which are optimally efficient, i.e. use perception tasks that are least susceptible to observer noise, and that concentrate on a small set of representative materials from which valid generalisations to all other situations can be made. This line of development has been followed for some time, especially by the SAM consortium, and could fruitfully be extended into the next five years.

The second way out is to replace the human observer by a computersimulated observer, i.e. to use automated assessment methods. Using automated methods presupposes that we know exactly how human listeners react to speech output. The development of objective methods is therefore necessarily subsequent to the development of human test methods. In those areas of auditory perception where sufficient, consolidated knowledge

has been assembled, attempts at computer-simulation can be launched even today, and, in fact, pilot studies have recently been undertaken that show the feasibility of objective testing in selected areas (see Section 12.2.4). The field will have to reach agreement on what further aspects of human perception, relevant to speech output assessment, have evolved to the point that computer-simulation of the human listener can realistically be undertaken. Once such areas have been identified, the next step will be to go ahead and implement them.

Candidates that present themselves for automated testing will be:

• Segmental quality: Here the problems are not unlike those that have to be solved in automatic speech recognition. In both cases automatic procedures have to be developed that accept/recognise speech as long as the stimulus is within certain prespecified ranges of tolerance, i.e. gets close enough to some prespecified ideal. So, common to both enterprises is the element that we have to be able to determine the (perceptual) distance between input and ideal. However, in automatic speech recognition the tolerances will have to be set as loosely as possible, so as to maximise the chances of deviating exemplars of sounds/words to be recognised. In speech output assessment, ideally, the same perceptual distance metrics can be used. Generally, the closer the speech output hits the ideal norms, the better its quality. Perceptual distance assessment will then become equivalent to quality assessment. Moreover, the results can be used not only at the level of overall assessment (e.g. in terms of some weighed mean perceptual distance), but also for diagnostic purposes: when specific sounds are poorly synthesised, they will show up as local maxima in the plots of perceptual distance as a function of time.

As a first approximation, such computer simulations should be tried for single speaker situations. That is to say, speech output should be compared only with human ideal speech produced by the same talker, pronouncing the same materials. Note that we assume that even allophone systems are based on a single model talker, since it is generally ill-advised to try and find average values over a larger group of speakers to control the synthesiser's parameters (Loman and Boves 1993, p. 159).

- *Temporal structure*: Using the same principles as above, the discrepancies in temporal organisation between system output speech and a human ideal can be determined (e.g. through Dynamic Time Warping).
- Melodic structure: Using the same principles as above, perceptual distance between speech output melody and melodic norm can be determined. Note that since there will always be (slight) differences in timing between speech output and ideal speech, both segmental and melodic assessment will necessarily involve temporal normalisation. The perceptual evaluation of the discrepancies between output speech and ideal should therefore proceed in – at least – two separate stages: first the penalty that is incurred by deviating durations will have to be determined, and only then can we meaningfully consider the penalty for deviating segmental quality (likewise for melodic structure).

We advocate a two-pronged approach here. The field should concentrate on developing optimally efficient tests involving human listeners, and at the same time begin to work on the development of perceptual distance estimation procedures that can be used later in automated assessment.

12.6.2.2 Predicting functional behaviour from judgment testing

There is a paradox involved in the choice between judgment tasks and functional tests. On the one hand, it could well be argued that a speech output system is adequate if a representative user group judges the system to be adequate for its purpose. Why should the field go to more trouble to improve the system's quality if the users profess to be satisfied? On the other hand, we can predict with near certainty that the users will not be able to estimate precisely the level of adequacy needed for the output system to function smoothly in a concrete application. The relationship between judgments and functional test scores has been studied in the context of segmental quality, but so far not in the field of prosodic quality testing. It would seem a point of immediate concern, therefore, to consider research into the interrelationship between judgments and functional test behaviour, with emphasis on prosodic quality. To what extent do orderings among competing speech output systems, as derived from judgment tests, correspond to orderings derived from functional tests? If we were able to predict functional test behaviour from judgment test scores, the latter, as a cheaper alternative for functional testing, could be used in all initial stages of speech output assessment. The use of functional testing would then typically be restricted to diagnostic testing.

12.6.2.3 Predicting global from analytic testing

Generally, one would expect the global quality of a speech output system to be a function of the quality of the various system components. One would like to be able to predict and quantify the overall ratings and global performance measures from the scores on the components through some form of regression analysis. Obviously, if system designers have only limited resources available, they should direct their efforts toward improving the quality of those aspects that contribute most (in terms of regression coefficients) to the overall assessment of his system. We suggest that research be undertaken in order to address this type of question.

12.6.2.4 Predicting field performance from laboratory testing

There is general agreement that laboratory tests such as are available today do not allow a useful prediction of how well a speech output system will perform in a concrete application. A short-term recommendation is, therefore, to develop a field-test generator, along the same lines as the successful test

generators for laboratory intelligibility tests (such as the CLID and SUS tests developed by the SAM consortium). The field-test generator should enable the fast compilation of test materials and adequate simulation of a range of application conditions. For this purpose, an adequate cross-section of applications for speech output has to be inventoried and parametrised along such dimensions as (1) type of users (non-cooperative, children, elderly people, non-native language users), (2) specific aspects of the situation in terms of, for instance, noise, reverberation, telephone channel, and (3) secondary tasks. An integrated software package, PMT (Parametric Test Manager) has recently been made available (by the Electrical Engineering and Acoustics Department of the University of Bochum, Germany) that contains some of the features proposed here:

- 1. data structure allows mixing stimuli with external audio and video signals, with audiovisual feature links,
- 2. signal editing capabilities,
- 3. contains query language for post hoc data analysis, and an interface with statistical packages,
- 4. allows cross-comparisons with alternative synthesisers and languages,
- 5. allows rhyme tests with both closed and open formats.

On a longer-term basis we advocate a more fundamental solution to the problem of field testing. Ideally, of course, one should not have to go to field every time a new application presents itself. Rather, one would like to be able to predict accurately, on the basis of available results of standard laboratory tests (e.g. intelligibility scores and prosodic adequacy profiles) how a speech output system would perform in a concrete field situation. For this to be the case, it will be necessary to have a valid analysis of the field tasks that have to be accomplished. A task profile will have to be drawn up that analyses the demands that carrying out the task (including and excluding listening to speech output) makes on the user, such as attentional load of the primary task, environmental noise, negative influence of fatigue and boredom, physical strain, etc. Accomplishing this type of prediction calls for cooperation between experts in the field of speech quality assessment experts and human factors studies. We recommend exploratory studies along the lines suggested above, based on quantitative task analyses of a few selected applications.

12.6.3 Linguistic testing: Creating test environments for linguistic interfaces

Generally, we feel that the development and testing of the higher-order linguistic modules of speech output systems should be left to language technology (NLP) experts (see also the results of the EAGLES Working Groups on Corpora, Machine Readable Lexicons, Formalisms, and Evaluation). A reasonable division of work would be for speech technology to deal with the mainly word-level linguistic modules that are specific to TTS applications, i.e. text preprocessing and grapheme-phoneme conversion (including stress position, accent placement and boundary marking). Other linguistic tasks, such as morphological analysis and syntactic parsing, are to a large extent common to other branches of linguistic engineering (e.g. grammar checking, automatic translation), with much more resources and manpower available. However, even if this division of work could be effected, one would like to see attempts made towards early separation of consequential vs. inconsequential errors in word and sentence parsers. Consequential symbolic errors will audibly affect the (quality of the) acoustic output, whereas inconsequential errors are not reflected at the audio level. This means that part of speech output testing should still be concerned with the higher-order linguistic modules.

We would advocate a more detailed analysis of the various tasks a text preprocessor has to perform, focussing on those classes of difficulties that crop up in any (European) language. Procedures should be devised that automatically extract representative items from large collections of recent text (newspapers) in each of the relevant error categories, so that multilingual tests can be set up efficiently. Once the test materials have been selected, the correct solutions to, for instance, abbreviation expansion problems can be extracted from existing databases, or when missing there, will have to be entered manually.

A short-term recommendation is to develop multilingual machine-readable pronouncing dictionaries at the single word level which list permissible variations. Comparisons of algorithmic output with the model transcriptions requires the development of adequate string alignment procedures. Moreover, not all discrepancies found contribute equally to the overall evaluation. Distance metrics should be developed that allow us to express the differences between two segmentally different phonemic transcriptions in terms of meaningful perceptual distance. Recent work done by Cucchiarini (1993) could serve as a starting point.

The correctness of most symbolic output can only be determined on the basis of connected text at the sentence level. What is dearly needed, therefore, is the availability of large, multilingual text corpora with full phonemic annotation, including not only the permissible pronunciation(s) of the words, including the effects of assimilation across word boundaries and stress shifts, but also the indication of accent positions (and degrees of accent), prosodic boundaries (with break indices of various strengths), and some intonation transcription. Moreover, since these corpora will also have to be used for testing morphological and syntactic parsing, hierarchical word and sentence structure should be indicated; or at least provisions should be made for linguists to enter this type of information at a later stage, resulting in a

hierarchically tagged text corpus or tree bank.

The development of corpora of this type is best left to the text corpus experts. We refer to the relevant chapters on database development (Chapters 3, 4, 5) for a discussion of corpus-related matters.

We recommend the development of procedures for strictly modular testing of linguistic interfaces. This means that test materials have to be made available that are specific to each individual module in the linguistic interface. Each module should be given correct input strings, and the correct output string(s) for only the module at hand should be provided. Only in this way can the problem of percolating and compounding of errors made by earlier modules be eliminated. Obviously, such procedures can only be effective if the databases referred to in the previous paragraph contain representations of the correct strings at each of the levels addressed by the various modules.

With the availability of cheap mass memory, the need for highly intricate rule-based linguistic interfaces is less strongly felt than some years before. Rather than computing the phonological code that is to be fed to the acoustic modules, the correct code is simply looked up in large lexicons included in the speech output system. If this trend continues, the emphasis of our research efforts will shift from rule development (and testing) to collecting databases. Database collection and annotation will take place regardless of the direction that the field takes in this matter. If choices have to be made, money is spent most safely on the development of corpora, but only if a multilingual notation format can be found that can be used for the transcription of segments and prosodic features of all languages dealt with. Although less important at the isolated word level, it still remain necessary to test grapheme-phoneme conversion. The output of post-lexical rules (changing the pronunciation of words in connected speech, e.g. through assimilation). Also, testing grapheme-phoneme conversion will remain applicable in the development of cheap speech output systems (such as MUL-TIVOX and APOLLO), which do not access large lexicons or perform sophisticated linguistic analyses of the input text.

12.6.4 Acoustic testing: Developments for the near future

12.6.4.1 Segmental quality testing

With a few provisos (see below) there is general consensus that the procedures for testing segmental quality of speech output systems are more or less fully developed (cf. Section 12.5.2.1.2 under DRT /MRT, CLID and SUS Tests). Under the auspices of the SAM consortium, efficient test generators have been developed that enable the construction of a large variety of tests that allow quick standardised administration and data analysis of consonant and vowel intelligibility scores, both for isolated word intelligibility and for intelligibility of words in (semantically unpredictable) context. These tools will be very useful in the testing of even the latest generation of parametric synthesisers. However, the upcoming generation of waveform synthesisers (PSOLA based) will have segmental quality that will be hard to discriminate from human speech. Though it may be possible to further refine the discriminatory power of our test procedures, one may well wonder what purpose would be served by such endeavors. A reasonable alternative view would be to consider the quality of waveform concatenation speech output equivalent to the human ideal (if indeed the test shows that no intelligibility difference remains) and leave the matter at that.

A short-term recommendation that should be made, concerns the quality of segments in unstressed syllables. It has correctly been pointed out by, for instance, Van Santen (1993) that most segmental quality tests consider monosyllabic words only (or "minisyllabic" words for languages without lexical monosyllables). There is a risk involved here that insufficient attention is being paid to the quality of unstressed syllables in longer words. The same, of course, holds true of the quality assessment of (unstressable) function words. Unstressed syllables are generally reduced in human speech, and synthesis-by-rule systems have often neglected to carefully model the reduction processes. In concatenative synthesis, the problem of unstressed syllables can be solved by enlarging the set of normally unreduced acoustic building blocks with a parallel set of reduced building blocks (cf. Drullman and Collier 1993). The testing problem that crops up in this connection presents an important perceptual question addressing the interaction between segmental and prosodic quality: if unstressed syllables are overarticulated, as would be the case when the reduction processes are not adequately modelled in our synthesis, does the resulting speech output get more intelligible, or does its intelligibility deteriorate? One might predict that, although the identifiability of each individual segment may decrease when reduction is truthfully mimicked, the overall intelligibility, in terms of word scores, will increase, reasoning that the rhythmic structure of words showing natural gradation of strong and weak syllables might be more important to word identifiability than optimal identifiability of each individual phoneme.

On a more general note, we suggest that serious attention be paid to differences in the contribution made to the overall intelligibility of words by the various constituent segments. It is important that we learn to what extent word intelligibility depends on identifying vowels versus consonants, in stressed versus unstressed syllables, in onset, medial, and final position, in short and longer words. Psycholinguistic studies on auditory word recognition have shown that, indeed, stressed segments – because of their greater inherent loudness and duration – have a better chance of contributing to the recognition process, as do segments early in the word. Ideally, we would
like to be able to predict the intelligibility of an arbitrary selection of words from the lexicon of a language, just by looking at the identification scores of the constituent vowels and consonants in unpredictable words (i.e. segment strings that are phonotactically legal and may be lexical words or nonsense strings).

12.6.4.2 Prosodic quality testing

With the advent of high-quality segmental speech output (Section 12.1.2) a shift from segmental quality testing to prosody quality testing seems imminent. It is obvious that there is still a long way ahead of us before the evaluation of prosody will get full coverage. What is needed is a careful taxonomy of prosodic functions at all linguistic and pragmatic levels (see also Chapter 5). We suggest, therefore, that the first priority should be for linguists to chart out all the prosodic functions relevant to human-machine communication. We need to know not only what functions are fulfilled by prosody, but also what the communicative importance of each specific function is (if any). Once a reasonably complete view of relevant prosodic functions has been obtained, attempts should be made at defining adequate tests in order to determine to what extent each function is expressed by the speech output system.

It will be difficult to separate the evaluation of prosodic forms from their communicative functions, and perhaps such a dissociation is not even necessary. It seems reasonable to assume that a prosodic feature fulfills its communicative function better as its formal properties are closer to the human model. If this relationship holds, we would not have to test the formal adequacy of speech timing and melody rules in abstraction from their communicative functions. Once we know the communicative function of each formal prosodic distinction, the prosodic quality of speech output systems can be measured by the effectiveness with which each of the communicative functions is signalled to the human listener. For these reasons we suggested that functional testing of prosody be given priority. Whatever audible flaws remain after the communicative functions have been shown to be signalled as effectively as in human speech will have to be addressed in a later stage, using judgment tasks.

We recommend therefore that the emphasis should be on the *functions* of prosody, rather than on the details of prosodic *form*. Our point of departure, for the time being, is that the formal aspects of prosody cannot be too far off the mark if the prosodic functions are all adequately fulfilled. This should not be interpreted in the sense that we consider the details of prosodic form (such as exact pitch movements and timing) unimportant. In fact, there is every reason to believe that prosodic functions such as accentuation are only adequately expressed by language–specific pitch movements which are very narrowly defined (in terms of direction, excursion size, and segmen-

tal alignment.) In this context it seems obvious that adequate prosodic functioning can only be guaranteed if speech output systems are capable of synthesising not only binary accent or boundary distinctions but also more subtle degrees of contrast within such categories. For instance, the adequacy of prosodic boundary markings should be tested at least at four levels of depth: strong and weaker boundaries within the sentence, as well as sentence and paragraph boundaries, which are signalled in parallel by melody, temporal organisation, and (possibly even) intensity.

Generally, we believe that the identification of prosodic functions to be tested (including the expression of emotion) presents a greater problem than devising tests to determine the functional adequacy of prosody once a particular function has been identified. Still, choices will have to made as to what particular test methodology to adopt. We propose that a pilot study be initiated to examine the pros and cons of the various tests used in the experimental phonetic and psycholinguistic literature (as outlined in Section 12.5.2.3.3) that seem relevant to this matter.

As a consequence of claiming priority for prosodic functions, the development of (multilingual) prosodic form tests (and test generators) should be postponed until some later stage.

12.6.4.3 Voice characteristics testing

It would appear that the evaluation of voice quality is going to be a matter of increasing concern. Developers of personalised voice speech output will need test procedures in order to determine how convincingly their systems mimic the quality of the model's voice. Simple same-different testing (*Is it Ella? Or is it Memorex?*) will not do, since developers will need the evaluation as a diagnostic tool. We suggest that a test tool be developed that enables the efficient drawing up of voice quality profiles (cf. Section 12.5.2.4.2).

Apart from the development of personalised voice synthesis, the voice quality of general purpose speech output systems will get a lot more attention in the coming decade. With the improvement of segmental, and to a lesser extent, prosodic quality of speech output, the need for more natural and pleasant voice quality will be strongly felt. It will be a concern for the evaluation field to develop test procedures in order to determine the appropriateness of voice quality for speech output in general and for specific applications (e.g. alert messages).

12.6.4.4 Overall quality testing

Now that the quality of speech output systems is getting closer to that of human speech, assessment should concentrate on other aspects of quality testing than linguistic functions. Synthetic speech may be virtually equivalent to human speech in all aspects, and still be lacking in certain subtle qualities. This aspect of speech output testing should be considered in a spe-

cial study, looking at the effects of listening to synthetic speech in terms of fatigue and allocation of attention to secondary tasks (cf. Section 12.4.1.1). The development of efficient multilingual test generators addressing this aspect would be a welcome addition to our repertoire.

12.7 Conclusion: summary of test descriptions

A limited number of tests, older as well as more recent ones, are described in detail below. The first 8 pertain to intelligibility tests, ordered from tests based on meaningless items of a few phonemes to tests based on meaningful sentences, 9 and 10 describe two prosody tests, and, finally, 11 and 12 are related to two overall quality tests. Tests were included that have been frequently used, or were developed specifically, for evaluating synthesis output. Four tests (SAM Standard Segmental Test, Cluster Identification Test, Semantically Unpredictable Sentences, and the SAM Overall Quality Test) were developed within the ESPRIT-SAM (Speech Assessment Methods) project (1989–1992). With the exception of CLID, software for these tests is available in SOAP (Speech Output Assessment Package) (cf. Howard-Jones 1992a). SOAP also includes programs to add noise to speech files, to automatically generate test sequences, to randomise test sequences, and to equalise speech levels. Test descriptions are given below; further current information can easily be found by standard Internet search methods.

12.7.1 SAM Standard Segmental Test

Status	Completely developed test with software (SOAP) for item construction, test administration, and data analysis (cf. Howard-Jones 1992a, Chapter 4).
Goal	Diagnostic and comparative evaluation of the intelligibil- ity of single initial, medial, and final consonants.
Languages	In principle any language. The only thing needed are lists of permissible consonants in initial, medial, and final po- sitions.
Items	Open vocabulary, mostly meaningless but also (by chance) meaningful, of the structure CV, VC, and VCV, comprising all phonotactically permissible combinations of initial, medial, and final consonants and three point vowels, e.g. /i/, /u/, and /a/. Examples: <i>pa</i> , <i>ap</i> , <i>apa</i> .
Procedure	Open response identification, preferably using unambiguous notation.
Time	With 20 consonants in a particular word position and 3 vowel contexts, a stimulus list contains 60 items. With an interstimulus interval of 4 sec this amounts to about 15 min for three lists. With two randomly ordered lists for each structure (recommended) time is doubled to 30 min per synthesiser.
Analysis	SOAP contains a pre-processing module to "clean" the responses. In addition it offers five scoring methods: (1) consonant confusion matrix, (2) percentage correct per consonant, (3) percentage correct per consonant per stimulus, (4) percentage correct per self-defined group of consonants, (5) group confusion matrix.

12.7.2 CLuster IDentification Test (CLID)

Status	Completely developed architecture with software for gen- erating tests to evaluate intelligibility at the word level: item construction, test administration, and data analysis (cf. Jekosch 1992).
Goal	Diagnostic and comparative evaluation of the intelligibil- ity of sequences of one or more consonants ("consonant clus- ters") and sequences of one or more vowels ("vowel clus- ters").
Languages	In principle any language, as long as matrices with permis- sible phoneme and phoneme combinations are available.
Items	Mostly meaningless but also (by chance) meaningful, open vocabulary taking into account phonotactic constraints; combinatory matrices are available for English, French, German, and Italian. The frequency of the syllable struc- tures (e.g. CCVCC, VCC) and clusters to be included in the test material can be defined. Examples: <i>storp</i> and <i>ast</i> .
Procedure	Open response identification using unambiguous notation or normal spelling.
Time	With 900 stimulus words (presented by Jekosch (1992), about 2 hours per synthesiser. However, subsets of stimuli can also be defined.
Analysis	The SOAP scoring module allows, among other things, the calculation of percentages correct of all clusters, of initial, medial, and final clusters separately, and of com- plete items. Confusion matrices are generated.

12.7.3 The Bellcore Test

Status	Completely developed test.	
Goal	Diagnostic and comparative evaluation of the intelligibil- ity of sequences of one or more consonants in initial and final word position.	
Languages	American English (cf. Spiegel et al. 1990).	
Items	Fixed set of 312 CVC-items, containing two tokens of nearly all onsets (N=62), nearly all plain offsets (N=63), and a subset of 31 (out of 111) offsets with phonetic af- fixes. Of each pair of tokens one is presented in a meaning- ful word, the other in a meaningless, but phonotactically legal word. Examples: meaningful versus meaningless on- set <i>swan</i> versus <i>swog</i> , plain offset <i>warmth</i> versus <i>dorth</i> , offset with affix <i>dropped</i> versus <i>globbed</i> .	
Procedure	Open response identification in normal spelling.	
Time	With all 312 items and an interstimulus interval of 6 sec (recommended), about 40 min per synthesiser.	
Analysis	Manually. Intelligibility is expressed as the percentage correct onsets and offsets or the percentage of cases where the phonemes occurring in the response correspond with a phoneme in the stimulus.	

12.7.4 Diagnostic Rhyme Test (DRT)

Status	Completely developed test.	
Goal	Diagnostic and comparative evaluation of the intelligibil- ity of single initial consonants.	
Languages	Originally English (cf. Voiers et al. 1975; Voiers 1983), adapted (at least) to Dutch (Steeneken 1982), and French (Peckels and Rossi 1973).	
Items	Fixed set of 192 meaningful CVC-words. Examples: <i>dune</i> and <i>tune</i> . The Dutch version has 384 stimulus words from which different lists of 192 stimulus words can be composed.	
Procedure	Closed response identification among two meaningful al- ternatives differing by a single phonetic feature in the ini- tial consonant. Both alternatives are presented as a stimu- lus, so both <i>dune</i> and <i>tune</i> are presented with the response alternatives <i>dune</i> and <i>tune</i> . Six contrasts are represented, namely voicing, nasality, sustention, sibilation, graveness, and compactness. Each contrast is included 32 times in the test, combined with 8 different vowels.	
Time	With an interstimulus interval of 3 sec, about 15 min per synthesiser.	
Analysis	Intelligibility is expressed as the percentage correct ini- tial consonants per contrast and for all initial consonants together. The percentages are corrected for chance.	

12.7.5 Modified Rhyme Test (MRT)

Status	Completely developed test.		
Goal	Diagnostic and comparative evaluation of the intelligibil- ity of single initial and final consonants.		
Languages	Originally American English (cf. House et al. 1965), adapted to German (cf. Sotscheck 1982).		
Items	Fixed set of (originally) 300 meaningful CVC words, con- sisting of six lists, each with the first 25 items varying in initial consonant and the final 25 items varying in fi- nal consonant. Later variants (cf. Logan et al. 1985) had mixed instead of blocked presentation.		
Procedure	Closed response identification among six meaningful al- ternatives, differing either in the initial or the final conso- nant. An example of a response set for a final consonant is <i>peas, peak, peal, peace, peach</i> , and <i>peat</i> . There is also a variant with open response identification among all mean- ingful CVC words known to the subject (cf. Logan et al. 1985).		
Time	With an interstimulus interval of 4 sec, about 25 min per synthesiser.		
Analysis	Intelligibility is expressed as the percentage correct initial and final consonants, or overall percentage correct conso- nants.		

12.7.6 Haskins Syntactic Sentences

Status	Completely developed test	
Goal	Comparative evaluation of sentence intelligibility, min- imising the effect of contextual cues.	
Languages	American English (cf. Nye and Gaitenby 1974).	
Items	Fixed set of 100 semantically unpredictable sentences of the form <i>The Adjective Noun1 Verb the Noun2</i> composed of high-frequency words. Example: <i>The old farm cost the</i> <i>blood.</i>	
Procedure	Open response identification in normal spelling.	
Time	With an interstimulus interval of 15 sec, about 30 min per synthesiser.	
Analysis	Manually, intelligibility is expressed in terms of the per- centage correct sentences or keywords.	

12.7.7	SAM S	emantically	Unpredictable	Sentences ((SUS))

Status	Completely developed test with software (SOAP) for item		
	construction, test administration, and data analysis (cf.		
	Howard-Jones 1992a, Chapter 5).		
Goal	Comparative evaluation of contence intelligibility min		
Guai	Comparative evaluation of sentence intemplointy, inin-		
	imising the effect of contextual cues.		
Languages	Dutch, English, French, German, Italian, Swedish (can be		
	extended).		
	, ,		
Itoms	Short computically unpredictable conteneos of five differ		
1001115	Short, semantically unpredictable sentences of five differ-		
	ent, common syntactic structures with words randomly se-		
	lected from lexicons with frequent, "mini-syllabic" words		
	(smallest words available in a given category):		
	1 Subject – Verb – Adverbial		
	og The table walked through the blue truth		
	2. Subject – Verb – Direct object,		
	e.g. The strong way drank the day		
	3. Adverbial – Transitive verb – Direct object (impera-		
	tive),		
	e.g. Never draw the house and the fact		
	4. Q-word – Transitive verb – Subject – Direct object,		
	e.g. How does the day love the bright word?		
	5 Subject – Verb – Complex direct object		
	or The place closed the fish that lived		
	Fifty sentences (10 per structure) are recommended per		
	synthesiser.		
Procedure	Open response identification in normal spelling.		
Timo	With 50 sontoneos and an interstimulus interval of 15 sec.		
TIME	with 50 sentences and an interstinuous interval of 15 sec		
	(recommended), about 15 min per syntnesiser.		
Analysis	Manually. Intelligibility is expressed in terms of the per-		
	centage correct sentences.		

12.7.8 Harvard Psychoacoustic Sentences

Ct - t	Conselstate developed test
Status	Completely developed test.
Coal	Comparative evaluation of contones intelligibility
Goal	Comparative evaluation of sentence intelligibility.
Languages	American English (cf. Egan 1948) French (cf. Combescure
Languages	1001) ID the (CD Interest of 1070)
	1981), and Dutch (cf. Plomp and Mimpen 1979).
Itoms	Fixed set of 100 meaningful syntactically varied phonet-
1001115	rixed set of 100 meaningful, syntaetically varied, phonet-
	ically balanced sentences.
	Example: Add salt before you fry the egg.
Drogoduro	Open response identification in normal spelling
Inforentie	Open response identification in normal spennig.
Time	With an interstimulus interval of 15 sec. about 30 min per
	sunthesiser
	synonesiser.
Analysis	Manually. Intelligibility is expressed in terms of the per-
J	aontara aonast kouworda
	centage correct keywords.

12.7.9 SAM Prosodic Form Test

Status	Proposal (cf. Grice et al. 1992a; Howard-Jones 1992a, Chapter 6).	
Goal	Diagnostic and comparative evaluation of parameters used to characterise specific intonation contours (e.g. falling and rising slope), regardless of the context in which they appear.	
Languages	English and Italian.	
Items	Meaningful words of varying segmental complexity in terms of phonological vowel length and syllable structure, e.g. (1) sketched, rain, moralise, salvageable (English) (2) re, basta, medico, nominano (Italian). Several items are included for each level of complexity, with accurate stress placement. Items are placed in a short, neutral carrier phrase, e.g. The word, or Dico Each item should be generated with the contours the system can produce (e.g. those triggered by a period ".", question mark "?", excla- mation mark "!", or comma ",") in addition to a reference monotone.	
Procedure	Rating of naturalness by naive subjects using magnitude estimation (see Section 12.3.2).	
Time	With 3 contours, 4 levels of segmental complexity, 5 items in each level, 4 repetitions of each token and an interstim- ulus interval of 5 sec, about 30 min per synthesiser.	
Analysis	Automatic calculation of the geometric mean of the re- sponses per synthesiser, segmental complexity, and con- tour.	

560 Part III: Spoken language system assessment

12.7.10 SAM Prosodic Function Test

Status	Proposal (cf. Grice et al. 1992b; Howard-Jones 1992a, Chapter 6).	
Goal	Diagnostic and comparative evaluation of the appropri- ateness of intonation contours for use in interactive com- munication contexts.	
Languages	English, but can be applied without effort to other lan- guages.	
Items	Two-part, human–machine dialogue excerpts.	
Procedure	Rating by naive subjects of the appropriateness of "pronunciation" using magnitude estimation (see Section 12.3.2). Example: Human: I'd like to reserve a flight to Paris on Monday morning.Synthesiser: Are you travelling from London? There is a choice between two protocols: The transcript of the excerpt appears on the screen, then (type A) the text produced by the synthesiser is played out to be rated with respect to appropriateness or (type B) both the text produced by the human and the synthesiser is played out, the latter to be rated. It is recommended to present fifteen exchanges per algorithm and to include test materials based on a hand-annotated intonation version as a reference.	
Time	Depending on the type of stimuli and the protocol.	
Analysis	Automatic calculation of the geometric mean of the re- sponses. Because of possible effects of differences in seg- mental quality between the dialogues on the ratings, it is advised to calculate the ratio between the text-to-speech version and the hand-annotated version.	

12.7.11 SAM Overall Quality Test

Status	Completely developed software (in SOAP) allowing the use of the magnitude and categorical estimation scaling methods. Two variants are recommended: 20-point cate- gorical estimation without reference (for test internal com- parison) and magnitude estimation by line length, with imaginary ideal speech as a reference (for test external comparison) (cf. Howard-Jones 1992a, Chapter 7).
Goal	Comparative evaluation of overall quality aspects, par- ticularly acceptability, intelligibility, and naturalness, for longer stretches of speech.
Languages	In principle applicable to any language as long as suitable stimulus material is available.
Items	Eight lists of 20 meaningful sentences of varying syntac- tic structures and length. For the rating of intelligibility and naturalness, speech material is available for Dutch, English, French, German, Italian, and Swedish. One list is sufficient for the evaluation of a synthesiser. Examples: I realise you're having supply problems, but this is rather excessive and I need to arrive by 10.30 a.m. on Saturday.
Procedure	Each aspect of speech is rated by a different group of sub- jects (minimally ten). When rating acceptability, it is rec- ommended that application specific speech materials are presented to (prospective) users. The ratings are based on two sentences each time.
Time	With 160 sentences and a 5 sec interstimulus interval, the rating of one scale takes about 20 min.
Analysis	Automatic.

12.7.12 ITU-T Overall Quality Test

Status	Proposal (cf. ITU-T 1993).
Goal	Comparative evaluation of overall quality aspects for longer stretches of speech.
Languages	In principle applicable to any language as long as suitable stimulus material is available.
Items	Speech samples of between 10 and 30 sec, adapted to the application. Example: Miss Robert, the running shoes Adidas Edberg Pro Club, colour: white, size: 11, reference: $501-97-52$, price: 319 francs, will be delivered to you in 1 week (mail order shopping). It is recommended that a (degraded) human reference is included
Procedure	Rating of (a subset of) eight categorical estimation scales (see below)
Time	With 4 test items per system, testing 4 synthesis systems with 3 reference conditions (i.e. 7 different sources) takes about one hour for one group of subjects, including in- structions to subjects and training session.
Analysis	Histograms and/or cumulative distributions of the ratings per scale and mean ratings. Little attention is paid to the answers related to content.

The eight categorial estimation scales:

- 1. Acceptance (Do you think that this voice could be used for such an information service by telephone?) 1: yes, 2: no.
- Overall impression (How do you rate the quality of the sound of what you have just heard?) 1: excellent, 2: good, 3: fair, 4: poor, 5: bad.
- 3. Listening effort (*How would you describe the effort you were required to make in order to understand the message?*) 1: complete relaxation possible, no effort required, 2: attention necessary, no appreciable effort required, 3: moderate effort required, 4: effort required, 5: no meaning understood with any feasible effort.
- 4. Comprehension problems (*Did you find certain words hard to understand?*)
 1: never, 2: rarely, 3: occasionally, 4: often, 5: all of the time.
- 5. Articulation (*Were the sounds distinguishable?*) 1: yes, very clear, 2: yes, clear enough, 3: fairly clear, 4: no, not very clear, 5: no, not at all.
- 6. Pronunciation (*Did you notice any anomalies in pronunciation?*) 1: no,
 2: yes, but not annoying, 3: yes, annoying, 4: yes, very annoying.
- 7. Speaking rate (What do you think of the average speed of delivery?)

much faster than preferred, 2: faster than preferred, 3: preferred,
 slower than preferred, 5: much slower than preferred.

8. Voice pleasantness (How would you describe the vo ice?) 1: very pleasant,
2: pleasant, 3: fair, 4: unpleasant, 5: very unpleasant.

Subjects hear each test item twice. After the first presentation they answer questions related to the content of the test items (8 sec), after the second time they rate the scales (about 20 sec).

13 Assessment of interactive systems

13.1 Introduction

13.1.1 About this chapter

$13.1.1.1 \quad {\rm Purpose}$

The purpose of this chapter is to present in accessible form a set of recommendations for the specification, design and assessment of interactive dialogue systems in which spoken language plays a significant part. The chapter is designed with simplicity and clarity of presentation in mind; its primary focus is on practical recommendations which the reader can easily apply in dialogue system development and comparison.

Work in this area is still at a very early stage of development. The recommendations contained herein must therefore be regarded as no more than provisional, born of the experiences and informed intuitions of those who have contributed to the chapter so far. References to the supporting literature are provided where appropriate but the literature on interactive spoken language systems is still rather sparse.

13.1.1.2 Scope

The technical focus of this chapter is interactive dialogue systems, defined most generally as computer systems with which humans interact on a turnby-turn basis. Our attention will thus be confined to those interactive systems in which "natural" language plays an important part in the communication process. Indeed, the recommendations are principally designed to apply to those systems in which natural language supports the majority, or the totality of the communication. More than this, we will focus particularly on those systems in which spoken natural language is the primary means of communication. Because there are many points of similarity between interactive dialogue systems based on spoken and written language, the recommendations seek to crystalise lessons learned from the written language domain. Also, some recommendations arising directly from experience in the spoken language domain can be expected to be portable to the text domain. However, readers whose primary interest is in text-based interactive systems should bear in mind that any material contained in this chapter which is of relevance to their concerns will not be marked explicitly as such.

Spoken language dialogue is, in many ways, the most challenging of all the language technologies because even the most basic dialogue system needs to subsume significant parts of most of the major fields of language engineering. The primary object of this chapter is to address issues which particularly arise when the various component technologies (such as speech recognition, parsing, database management, linguistic generation, speech synthesis, etc.) are integrated in a single functional system. Readers would do well to consider whether they require more detailed information relating to any of these component technologies. If so, other chapters in this volume, and other volumes in the EAGLES series, will provide good entry points to these fields.

13.1.1.3 Terminology

In an interdisciplinary technical field such as interactive dialogue processing, the importance of mastering the terminology should not be underestimated. Technical terms are defined where they first appear in the text.

13.1.2 Reading guide

13.1.2.1 Different needs addressed

Users of this handbook are likely to have needs which fall into three distinct categories. First, end user organisations will wish to compare existing interactive dialogue systems in order to select the best solution for some particular purpose. Second, system engineers will wish to gauge the performance of existing systems for diagnostic purposes, in order to improve their performance. Third, system designers will wish to learn how to go about designing an interactive dialogue system from scratch. These three needs are addressed below.

13.1.2.1.1 Comparing existing systems

It is notoriously difficult to compare existing dialogue systems, as they really ought to be compared in exactly the same conditions. In fact, comparison depends on the degree of system integration. Dialogue managers (as distinct from dialogue systems) should ideally be evaluated independently of the speech technology (recogniser and synthesiser) and of the application domain. In fact, this is rarely possible: for instance, dialogue prediction and correction procedures are heavily dependent on the recogniser performance and its linguistic analysis components. A dialogue system is also rarely completely independent of the application domain or, at least, of a class of applications. Even for the same application, the interface might be different (for air-traffic control training, for instance, there exist different air-traffic simulators with different levels of complexity). Interfaces between the system and the speech technologies on the one hand, and the system and application on the other are not at present general-purpose. Adaptation is always necessary. Complete systems developed for the same application domain could however be assessed on corpora of similar complexity, corresponding to the same pre-defined scenarios, but as they have different

internal architectures, with different actual components (which need not coincide with abstract components), only a black box assessment might be envisaged.

This chapter aims to make these issues accessible to people who may lack extensive experience in speech and language technology and who wish to compare existing systems.

13.1.2.1.2 Improving existing systems

Improving existing systems may aim either to improve the system performance (overall or parts of the system), or to render the system more independent of either the speech technologies or the application.

- Improving system performance can be achieved by means of an iterative process described below, or by conducting independent tests of single components (parser, interpretation module, for instance) or parameters (using different prediction mechanisms).
- Rendering the system independent of the application (vocabulary, syntax, etc.) requires the development of specialised interfaces and specific tools which allow the user easily to change the semantic domain by describing the associated knowledge (vocabulary, constraints rules, etc.) in an interactive way. This requires that a knowledge compiler tool exists which can compile the data into a form usable by the dialogue system, and also that coherence verification procedures exist to assure the compatibility of the content of syntactico-semantic knowledge with the linguistic information in recogniser and synthesiser languistic analysis modules.
- Rendering the system independent of the speech technology helps to make it portable and not reliant on particular technologies. However, this may not serve the needs of improving system performance, as better performance is likely to be obtained if the system is tightly tuned to the recogniser performance, at the expense of its portability.

By outlining a framework for testing, respecifying and enhancing systems, this chapter provides a way into this complex problem.

13.1.2.1.3 Designing new systems

A background activity to designing new systems is to try to assess existing systems to the limit of their possibilities (maximum number of words, for instance), assigning limit values to their variables (vocabulary size, number of semantic frames or concepts, etc.). The results of the evaluation of systems which deal with similar tasks will also be of considerable relevance here. Besides, designing new systems assumes that several analyses have been done beforehand, based on the following procedures:

• Analyse the usual human interaction if such a model exists, preferably with real recorded dialogue.

- If necessary, take speech technology performance into account, and specify the common knowledge which will have to be shared (the linguistic information used in the recognition and analysis process, for instance).
- Specify the interface with the application (specific coded language used such as SQL, for instance), and the common knowledge which will have to be shared between the system and the application (which may be crucial when the application domain knowledge is evolving during the dialogue). This will help to design the task model.
- Define the knowledge to be included in the user's model (taking classes of users into account, if necessary).
- Define the dialogue strategies which are necessary (depth or shifts authorised, corrective and/or predictive procedures when necessary).

These, and other related tasks, are explained in this chapter in the context of system specification and design, along with some detailed procedures for progressing from an initial goal to a final working system (see also Chapter 2).

13.1.2.2 Section summary

It is important to understand exactly the nature of the technology with which this chapter is concerned, and to master the technical terms which will crop up again and again throughout the chapter. These needs are addressed in Section 13.2, on *Interactive dialogue systems*.

Interactive dialogue systems are highly complex systems, incorporating many different technologies. Section 13.3, *Specification and design*, reviews some of the approaches which have been adopted to the problem of specifying and designing such systems. This section concentrates on specifying the functionality of interactive dialogue systems. Detailed recommendations based on practical experiences of workers in the field are included.

Once an interactive dialogue system has been specified, designed and implemented, the task of assessing how well the system performs is a non-trivial task. Section 13.4, *Evaluation*, looks at what makes the problem difficult, describes a framework within which evaluation may take place, and suggests a core set of metrics which can be used for comparing different interactive dialogue systems.

13.2 Interactive dialogue systems

13.2.1 Definitions

An interactive dialogue system is constructed such that it enables and supports the communication between a human user and the service offered by the system. It is based on the integration of a set of modules, each of which handles a complex task. The modules are linked to each other and their interactions are controlled by a *kernel module* which has the overall task

of managing the dialogue. Seen from the *dialogue manager*, the application functions as an external module (e.g. a remotely functioning database) connected to a human user who may have a number of input and output devices at his disposal.

A dialogue manager may be able to handle several input and output devices in parallel: a user may interact with the dialogue system using multimodal input and output, and several input devices may be used in transferring the same message to the system, for example, DTMF (touch tones) instead of speech input.

Users communicate with the system in a number of transactions. A transaction consists of a number of exchanges, each of which consists of the input utterance (or a sequence of DTMF signals for a touch tone input device) or the corresponding system response (e.g. synthetic or canned speech or text on a screen). The attention of the interactive dialogue system changes in a sequence of turns.

A number of basic terms from interactive dialogue are introduced here:

- INTERACTION: Communication of information between two agents (parties), each of whom is capable of actions. A signal or a stimulus (for example, an utterance in a linguistic interaction) coming from one agent provokes a change in the internal state(s) of, and a response(s) from, the other agent. This process is symmetrical between both parties. Of special relevance here is the case where one of the agents is a human being and the other is an automatic system. Stimuli are understood in a broad sense, including multimodal stimuli (in which different media may be used). They may, for example, consist of a physical action (moving a pointer) or a linguistic act (uttering a sentence) or the coordination of both.
- SPOKEN LANGUAGE DIALOGUE: Also known as oral dialogue. A complete spoken verbal interaction between two parties (in the present case, a system and a human being), each of whom is capable of independent actions. A dialogue is composed of a sequence of steps which are related and build on each other. Dialogue systems are more sophisticated than question/answer systems, in which one agent may pose a succession of unrelated queries to the other agent.
- TASK-ORIENTED DIALOGUE: A dialogue concerning a specific subject, aiming at an explicit goal (such as resolving a problem or obtaining specific information). For example, dialogues concerned with obtaining travel information or booking theatre tickets are task-oriented.
- TRANSACTION: The part of a dialogue devoted to a single high-level task (for example, making a travel booking or checking a bank account balance). A transaction may be coextensive with a dialogue, or a dialogue may consist of more than one transaction.
- TURN: A stretch of speech, spoken by one party in a dialogue. A stretch of speech may contain several linguistic acts or actions. A dialogue consists of a sequence of turns produced alternately by each party. Turns are also known as utterances.

- EXCHANGE: A pair of contiguous and related turns, one spoken by each party in the dialogue.
- WIZARD-OF-OZ SIMULATION: Simulation of the behaviour of an interactive automaton by a human being. This can be done (i) by speaking to the user in a disguised or synthetic voice, (ii) by choosing and triggering system predefined responses, (iii) by manually modifying some parameters of the simulation system, or (iv) by using a person to simulate the integration of existing system components, i.e. a bionic Wizard-of-Oz simulation (see also Chapters 4 and 9).

Now that these basic terms have been defined, we shall consider how interactive dialogue systems compare with command systems, and shall review some issues relating to the different levels of interactive complexity to be found in dialogue systems.

13.2.1.1 Interactive dialogue systems compared to command systems

In command systems, the interaction is direct and deterministic: to one stimulus from one agent corresponds one unique response from the other agent, the response being independent of the state or context of each agent. For example, you press a key on a keyboard and the expected character appears on the screen. With command systems, the human has direct control over the machine. This form, not normally considered as a variety of human communication, is usually referred to as the *tool metaphor*.

A dialogue system can be considered as a kind of interface which performs communication between a human being and an application system, which may include several other systems. The dialogue system must process two kinds of information: that coming from the user and that coming from the task itself through specialised interfaces, one for the speech technologies, one for the application. One of the dialogue system's main activities is to maintain coherence between both. Therefore, the connection between a human being's action (a natural language utterance, for instance) and the response of the system is not direct: the dialogue system must achieve a number of internal actions in order to give a response which is not unique but depends on the internal state of the system and on the context of the interaction. This form of communication is referred to as the *agent metaphor* or the *advisor metaphor*.

13.2.1.2 Interactive complexity of dialogue systems

Dialogue systems include different comprehension levels relating to basic components: a recogniser, a parser, an interpretation module, a dialogue manager, a synthesiser, etc. Each of the modules requires associated knowledge databases (lexicons, rules and models concerning the language used, the system, the task, the user, the environment, the dialogue itself). Each of the models has both static and dynamic parts: the static part exists before

the dialogue begins, the dynamic part is built and modified during dialogue. One important component is the dialogue history which keeps track of the previous exchanges. The different modules and their associated knowledge bases allow the dialogue manager (or system) to perform internal actions including the following:

- Verify the coherence of the user's request with the system.
- Knowledge concerning the linguistic analysis and generation modules, the task, etc.
- Negotiate with the user.
- Resolve problems of reference (anaphoras, ellipses, etc.).
- Generate a relevant response.
- Draw reasonable inferences.
- Predict the user's most probable reaction.

The different comprehension levels involved (acoustic, phonetic, lexical, syntactic, semantico-pragmatic) may be addressed sequentially. Alternatively, information transfers may take place in parallel between different levels in a non-hierarchical fashion, depending on the dialogue situation.

The role and performance of the dialogue system are largely constrained by and therefore dependent on the performance of speech technologies (depending on the recogniser error rate and authorised vocabulary, or on the control parameters of the synthesiser, for instance). They are also greatly dependent on the task objectives and requirements.

Different interactive complexity levels in dialogue systems may be identified. These are described in the following sections.

13.2.1.2.1 Menu dialogue systems

The interaction is reduced to a question—answer user-interface. The dialogue model is merged into the task model from which it cannot be distinguished. Dialogues of this kind are often represented by branching tree structures. This category includes interactive voice response (IVR) systems, integrating tone signalling, isolated word recognition and word spotting techniques. The dialogue is strictly guided, leaving very little initiative to the user (system utterances may in some cases be interrupted by the user, for example). Several exchanges may be necessary to provoke one action or to obtain information from the system. This latter feature distinguishes these systems from pure voice control or command language systems in which there is no dialogue.

A question/answer system is a particular limiting case, as it may either be considered as a command system or as a marginal dialogue system: if one particular question always provokes the same response whatever the situation, then the system may be considered as a command system. But if asking the same question can provoke different responses (in menu-driven dialogue systems, for instance, it may depend on the current level in a tree structure), then the system can be called an interactive dialogue system.

13.2.1.2.2 Spoken language dialogue systems

The system possesses distinct and independent models for the task, for the user, for the system, and for the dialogue itself. The dialogue model takes context into account, using a particular knowledge base (a dialogue history), which is built during dialogue. Multiple types of references (anaphora, ellipses), may be processed. The system may be capable of reasoning, of error or incoherence detection and internal correction, and of anticipation and prediction.

13.2.1.2.3 Multimodal dialogue systems including speech

In this case, the complexity of the spoken language dialogue is compounded by the fact that the result of speech recognition has to be merged with other information delivered by other means of communication (media). The dialogue is itself dependent on the system model. Each piece of information delivered by a medium must be dated, as each medium does not process information in the same time, and the dialogue manager has to take event chronology into account.

The first category of systems (menu systems) is now used in several realworld application domains (enquiries about cinema programmes, travel timetables, bank accounts, etc.). Most applications deployed in the field work over the telephone and are used by the general public. Members of the two other categories are mostly still industrial and laboratory prototypes, which still impose a lot of constraints (such as a training phase, and a quiet environment) on the user. However, this position is steadily changing as more advanced interactive systems come to be deployed in the field.

13.3 Specification and design

The design of a specific application is a complicated task which in many cases involves the integration of a number of speech technology components. This may include a speaker independent, continuous speech recogniser, a parser which is able to perform error-recovery, a speech synthesiser, and an application, all being controlled in an integrated fashion by the dialogue manager. Interactive dialogue systems, however, have to interact especially with the human user, who may never have used such a system before.

In this section we consider the process which leads from the high level goal of creating a dialogue system to operate in some domain, to the development of a detailed design on which the implementation may be

based. This is a non-trivial task which practitioners new to the field should not underestimate.

Recommendation

Always specify an interactive dialogue system fully before implementing it.

Full specification is advisable for the following reasons:

- It gives a good indication of the scale of the task.
- It allows potential problems to be spotted early.
- It usually leads to cleaner and more modular code.
- It gives all parties with an interest in the system an opportunity to check that they agree on the proposed functionality of the system prior to implementation.
- Most importantly, it forces the task of system definition to be given the attention it deserves.

This recommendation is no different from good practice in any area of software engineering. However, interactive dialogue systems engineering is a young discipline, and it still suffers from an attitude which is common amongst software researchers, namely "implement first, design later". Experience suggests that dialogue design is a harder and more time consuming activity than dialogue implementation, and it is worth taking this recommendation seriously.

A specification may consist of the following sections:

- 1. An overview of the proposed system, describing in simple terms what the system is for, who will use it, etc.
- 2. A formal and explicit description of the proposed dialogue. This could be expressed, for example, as a state machine with a description of all reachable states in the dialogue, together with information on what actions should be performed in each state and how to decide which state should be the next. It could be portrayed graphically as a flowchart or a state transition network, or it could be represented as a dialogue grammar. Whatever representational choice is made, the key point is that the specification should be complete and unambiguous. Preferably, it should also be accessible to potential end users, so that their views on the specification can be elicited.
- 3. A description of the platform and general computing environment on which the dialogue system will be delivered.
- 4. Any other relevant technical material, such as communications protocols for interfacing with external databases.

There are currently no flowcharting standards for spoken language dialogue system specification. However, general purpose standards do exist which can be appropriated and modified where necessary for this purpose. For example, the SUNSTAR Project, and subsequently the Danish National Project on Spoken Language Dialogue Systems, uses DDL (Dialogue Description Language), whose graphical component is based on SDL (Specification and Description Language), which was standardised by CCITT (cf. Belina and Hogrefe 1988). It is basically a graphical language for describing state transition diagrams for event-driven systems. It has been heavily extended with new symbols, new meanings for existing symbols, new diagram types, and new ways of connecting symbols.

In the following sections we shall examine three approaches to dialogue system design which have been explored by researchers. Few groups have used only one of these approaches to the exclusion of the other two. As we shall see, each approach has different strengths and weaknesses. We shall call these approaches *design by intuition, design by observation* and *design by simulation*.

13.3.1 Design by intuition

Humans are natural conversationalists. As very young children we easily learn to follow and participate in linguistic exchanges. Even before we have mastered a single word we make an effort to engage in some kind of interactive exchange by making non-linguistic noises by turns with talking adults. Once acquired, natural language endows us with the ability to form complex ideas and discuss them with others. One of the most remarkable aspects of language is its reflexivity: we are able to use language to talk about language. Humans, then, are in principle authorities on language who, conveniently, have the capacity to articulate that knowledge by means of language. Perhaps the best place to look for data on spoken language dialogue is within, in our own expert intuitions.

This view is prevalent in theoretical linguistics. The most famous statement of the doctrine of the primacy of *native speaker intuitions* derived from linguistic *competence*, i.e. the underlying knowledge of language, as opposed to *observations* of data derived from actual linguistic *performance*, was offered by Noam Chomsky:

Linguistic theory is concerned with an ideal speaker-hearer in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance (Chomsky 1965, p. 3).

Chomsky's robust statement of mainstream linguistic thinking must be set in its historical context. Throughout the 1940s and 1950s linguistics, particularly North American linguistics, was dominated by empiricist structuralism (cf. Hockett 1958). The practice of linguists was to go out into the field

and collect large amounts of language data, i.e. they recorded and transcribed actual utterances. This data was examined using simple analysis techniques which as far as possible kept sophisticated intuitions about the nature of language outside of the process. The prevailing view was that intuitions were not to be trusted; they might turn out to be no more than noise distracting researchers from spotting the true – possibly counter-intuitive – regularities in the structure of language. For example, the structuralist approach to establishing the word classes in a language was to work with the minimal criterion of substitutability. If two words can appear in exactly the same sentential context, then they are strong candidates for membership of the same word class.

There were at least two problems with the structuralist approach. First, the number of sentences in a natural language is, practically speaking, infinite so that it is never possible to be sure that enough data has been acquired to motivate a generalisation. Second, structuralism provided no way of distinguishing noise from reliable data. For example, the following two utterances must be treated equally, in spite of the fact that the second is unacceptable to an extent which makes it a potential subject of comment:

- a. I used to be able to run all the way to the station.
- b. I used to could run all the way to the station.

The point here is not whether or not (b) is interpretable. Rather, it is that the very same speaker who uttered (b) also has the capacity to reject it as a "slip of the tongue", if necessary, i.e. as a failed attempt to utter (a). Empiricist structuralism closed its ears to the rich data source offered by explicit statements of native speaker intuitions.

The Chomskyan method in linguistics, which has largely replaced empiricist structuralism, treats intuitions as primary. Even though linguistic acceptability (i.e. grammaticality) judgements turn out to be graded rather than binary, the last thirty years of research has not come up with a better source of data than the intuitions of linguists themselves. This applies particularly to the analysis of dialogue, though powerful statistical methods have been introduced to supplement linguistic categorisations for smaller linguistic units, and increasingly extensive dialogue data, with dialogue act annotation and stochastic analysis, are being used to train *dialogue models* for spoken language systems.

So how does the experience of theoretical linguistics relate to that of spoken language system design? In fact, the relationship has been close so far. Most researchers in natural language processing (NLP) have a background or interest in linguistics. Perhaps as a result of this, the vast majority of NLP systems have been oriented towards competence rather than performance. This is how dialogue system designers have tended to proceed.

• First, a high-level goal is fixed. For example, the goal may be to provide

interactive spoken language access to a database.

- Next, the task is analysed in detail to establish its parameters, and the routes by which the task can be accomplished.
- The next step is to make the leap from an understanding of the task to a specification of the ways in which the task can be linguistically realised. The problem is tackled by introspection based on expert intuitions.

At each stage in each possible dialogue, the designer attempts to answer the question "What could happen next?" This can be answered at multiple levels. For example, an acceptable answer for a given point in a banking application dialogue could be "The user will ask for an account balance." This answer abstracts away from the surface realisation of utterances. It could be realised in many different ways, including the following.

- a. How much is in my account?
- b. What is my present balance?
- c. Can you give me a balance, please?
- d. What have I got at the moment?

Each branching point in a natural language dialogue offers very large amounts of variability. If the question "What could happen next?" is answered at the level of abstract task-oriented speech acts, there are likely to be several possible next moves with non-negligible probability. This modest branching is complicated by the fact that there are potentially very many different ways of linguistically realising each distinct abstract move type. In turn, each of these linguistic realisations may be acoustically realised in infinitely many different ways. The whole solution space cannot be explored by means of introspection alone, although carefully chosen subparts of it may be designed quite effectively on the basis of designers' intuitions. Empirical support for this claim comes from those linguists who have tried to construct computational grammars capable of analysing all the sentences in substantial corpora of real language use.

Most dialogue systems developed to date have side-stepped some of the problems inherent in interactive dialogue by ruling out mixed-initiative turntaking in dialogue. Whereas in natural conversation participants cooperate and collaborate in the unfolding of the talk, dialogues involving designed systems tend to require the user to follow a strictly predetermined fixed-initiative dialogue plan. In these menu-based systems, standard conventions are set aside and the system conveys to the user the non-negotiable rule of interaction, namely "I ask the questions; you answer them." The following example should convey how menu-based dialogues are typically constrained.

System: Thank you for calling the EuroWatch Weather Forcasting Service. For which country do you require a weather forecast?

```
User:
       The United Kingdom.
System: Scotland, England, Wales or Northern Ireland?
User: Scotland.
System: Do you want a weather report for today, a short term
       forecast for the next 48 hours, or a long range
       forecast for the next 7 days?
User:
       A short-term forecast please.
System: Here is the weather forecast for Scotland for the
        next 48 hours:
        It will begin wet in all regions and will grow
        progressively wetter throughout the next two days.
        Do you require any further information?
User
       No thank you.
System: Thank you for calling. Good bye.
```

In this example, the system asks closed questions. There may be some flexibility as to the exact wording of a user response. For example, the user's first utterance could reasonably have been any of the following (parentheses enclose elements which can be omitted): "((for) the) United Kingdom (please)", "((for) the) U.K. (please)", "(for) Britain (please)" or "(for) Great Britain (please)". However, the directness of the system's prompting effectively rules out the much wider range of utterances which could include examples such as "Oh hello. I was wondering if you could give me a weather forecast for the U.K. please." While it is reasonably straightforward to construct a linguistic model which covers most utterances that would be produced by cooperative users in the more constrained case, the same cannot be said of the case where the system asks open questions or allows user initiatives.

In designing a strictly system-led menu dialogue system, all that is required of the designer is to come up with some way – any way – of allowing each task in the application domain to be performed. This approach could be called "*a priori* design" – the system designer states in advance how the user is to be allowed to progress through the task to the goal. The designer does not look beyond existing intuitions about how best to structure dialogues in order to develop a working system. There are, of course, no guarantees that the chosen design will be ergonomically optimal, and it will be necessary to test the system with users in order to fine-tune the dialogue strategy.

However, once a decision has been taken to renounce fixed initiative, menubased dialogue in favour of freer, more natural spoken language dialogue, the use of intuitions alone in specification must be called into question for at least three reasons.

1. The space of possibilities is too large

For any non-trivial task there are simply too many alternatives at each turn transition place in talk. There is a strong possibility that any attempt to build a wide coverage dialogue system will be riddled with gaps in coverage for which the only defence on the part of the designer can be "I didn't think of those."

2. Intuitions about performance are unreliable

Recall that the respectable tradition surrounding the use of intuitions in theoretical linguistics is earthed in linguistic *competence*. When designing a practical dialogue system it is necessary to design for *performance*. There is no evidence to suggest that sophisticated native speakers of a language possess reliable intuitions about language use. It is a standing joke amongst theoretical linguists that people who hold strong views about what they take to be the abuse of language are capable of deploring certain "debased" forms of language while simultaneously employing these very forms in their deprecating pronouncements (The example "Prepositions are words you should never end sentences with" nicely caricatures this kind of statement). People are remarkably insensitive to the exact surface forms used in talk. We extract information from utterances and then proceed to forget the detail. This blindness to the detail of performance is so pronounced that it is not uncommon for speakers to have forgotten exactly how a sentence began before they have finished uttering it.

In producing utterances, speakers seldom end up saying exactly what they wanted to say. In listening to other people's utterances we filter them with expectations to such an extent that we are quite capable of hearing what we want to hear and not what has really been said.

Consider the case of non-lexical fillers in speech, such as *uhm*. If asked, most speakers would suggest either that these items have random distribution or that they are inserted when the speaker cannot think what to say next. In fact, research has shown that these items are used in a highly ordered way to structure talk and to aid the smooth transfer of turns between speakers. This is a fact of considerable consequence to system designers, but one which would not have emerged through introspection.

3. We have no intuitions to rely on anyway Intuition is a human capacity to make judgements based on internally stored knowledge. The problem inherent in trying to design a spoken language dialogue system based on intuitions is that we have no experience of how people will react when confronted with a conversational computer. It may be that they will react very much as they react to any conversational partner; but this is just a guess which may turn out to be mistaken.

In summary, design by intuition is a cheap, simple and effective approach for specifying and designing system-led, menu-style, interactive voice response systems which, implicitly or explicitly strongly limit the kind of language which the user may employ. It requires no special materials or resources other than those with which system designers are already naturally endowed. However, there are fundamental problems at the root of this method when the dialogue system in question allows less constrained natural language interaction. Though the design by intuition approach may be a useful compliment to other approaches (such as design by observation and design by simulation), it cannot be relied on as the primary specification/design methodology.

RECOMMENDATIONS

Use design by intuition as the primary specification/design methodology only for those applications in which the following conditions apply: (i) all tasks in the domain can be structured into a fixed sequence of steps, (ii) the system takes the initiative in all phases of dialogue, and (iii) the design of system prompts and the nature of the domain constrain the kind of language which may reasonably be employed by the user.

13.3.2 Design by observation

Until relatively recently there has been very little discussion of data in the natural language processing (NLP) literature (but see the results of the EA-GLES Working Group on written text corpora). While speech technologists concerned themselves with the task of of collecting, managing and exploiting speech corpora, computational linguists tended to work primarily with native speaker intuitions, though an increasing number of researchers practise corpus-based NLP. The lack of attention to real text data is now beginning to be addressed through a growing interest in the collection of corpora which are used to define the coverage of grammars and parsers. For example, the Linguistic Data Consortium established with U.S. Government funding aims to collect and distribute large quantities of computer-readable speech and text. This is just one of a number of initiatives around the world to collect really large corpora (counted in millions of words) from which to extract information about aspects of language use, a key area also addressed by the EAGLES initiative.

In the past, much of the data for speech recogniser training has been generated on demand by getting a number of people to read each of the words to be learned (or sentences containing them), perhaps several times. This approach, though yielding reasonably good results, has to some extent been discredited by work demonstrating the significant differences which exist between read and spontaneous speech (for example, see Soclof 1990). It is becoming increasingly clear that in order to obtain the best possible training data for speech recognition systems it is necessary to collect and analyse samples of real spontaneous speech. Needless to say, this is a much more costly and time-consuming exercise than simply getting a small group of cooperative subjects to read prepared scripts in the laboratory.

Corpora of spoken language have much to offer for the training training or corpus-tuning of speech recognition and speech synthesis systems, in particular for stochastic language models and Hidden Markov Model techniques in the decoder stages of speech recognition. Do they have a role to play in the process of designing spoken language dialogue systems?

The answer to this question must be a clear "yes". Many of the objections to using native speakers' intuitions as design data are addressed using by observational data. One of the problems with intuitions was that the space of possibilities in spoken language dialogue is extremely large: there are just too many different possibilities to allow the designer to explore them by introspection alone. What a corpus of dialogues offers is concrete evidence to give the system designer a strategy for handling the problem. Speech recognisers in general support a restricted finite vocabulary, bounded by the limitations of the current technology. Suppose that some speech recognition system is only capable of operating in real time if it has a lexicon of 100 words or less. The most reasonable way to decide which words to include in the lexicon (assuming that the user is not explicitly restricted by means of a menu) would be to select the 100 most frequently occurring words in the chosen domain of discourse. While people can perhaps make fairly good guesses at the two or three most frequently occurring words, no-one could discriminate reliably between the 100th and the 101st words on the basis of intuitions alone.

What is true in the case of lexical selection is also true for higher levels of the system. Suppose at a given point in a human-computer dialogue, a cooperative user could reasonably say almost anything. The only way to design a practical dialogue manager would be to equip the system to deal with the most likely cases and to provide it with a repertoire of general purpose recovery strategies to enable it to repair understanding failures and proceed in orderly fashion with the dialogue when unanticipated utterances are produced. The task of prioritising which cases to manage specifically and which to leave to general failure repair mechanisms is, perhaps, even more difficult than that of selecting which words to include in the recognition vocabulary. By observing human-human dialogues in the chosen application domain it is possible to base these difficult design decisions on a solid foundation of empirical fact, rather than on the shifting sand of mere conjecture.

So far, all that we have done is to argue for the proposition that *some* empirical data is better than none. Now we must pause to consider just how reliable human-human dialogues can be as data sources for dialogue system design.

Most languages have a "Standard" dialect which serves as the basis for NLP systems, sometimes to the exclusion of a significant number of speakers of the language. However, it should not be thought that (ignoring minor ideolectal differences between speakers) dialects are constant across all situations of use. On the contrary, each dialect encompasses a rich variety of different registers. A register is a variety of the language which is selected according to the context of use. So, for example, one might greet an old friend with the word Hi, a bank manager with the words *Good morning*, and a complete stranger with the words *How do you do?* Consider the linguistic differences to be found in a conversation about the weather with a small child or a conversation on a similar theme with a potential employer during

a job interview. The interactions are likely to differ markedly in lexical selection, grammatical structure, formality, intonation, and indeed almost every conceivable linguistic dimension.

Speakers are able to reason to some extent about how they might speak in completely new situations. For example, many people will have experienced exchanges with a massively aggressive or brutal interlocutor, most people have the experience of being the weaker partner in a hierarchical relationship, have had conversations with individuals of whom they are afraid, have been asked difficult questions in personal contexts or in examinations, have at some time lost face in conversation, and so on.

What no-one yet has any extensive experience of is engaging in fairly free natural language dialogue with an asocial artificial being. A very considerable part of human-human talk is taken up with interpersonal concerns. Indeed, a legitimate question which has been raised is whether or not the word "conversation" makes any sense in the context of human-computer dialogue, so intimately entwined are our notions of conversation and social interaction (cf. Button 1990). It is hard to know where to begin speculating how people might react when faced with a non-human dialogue partner. The one safe bet which can be placed is that, in the words of Jönsson and Dalbäck (1988) "talking to your computer is not like talking to your best friend".

A result of this is that we have no safe grounds for extrapolating a detailed specification of a spoken natural language dialogue system on the basis of a corpus of human-human dialogue, even if this corpus contains dialogues addressing the tasks foreseen for the planned interactive dialogue system in the target domain.

In summary, human-human dialogue data has considerable value for building an understanding of the domain and its component tasks. In the absence of other information, it can be used to construct initial vocabularies and language models for systems. However, time should be allowed for refining the vocabulary and language models to compensate for possible linguistic changes introduced by the non-personal nature of human-computer dialogues.

RECOMMENDATION

Where possible, use human-human dialogue data to build an understanding of the domain and its component tasks.

RECOMMENDATION

In the absence of simulation data, use human-human dialogue data to create vocabularies, language models, and dialogue automata, augmented where necessary by careful use of linguistic intuitions. Before leaving the topic of observational data, it is worth mentioning a technique which has recently emerged, primarily for the purpose of collecting speech data for training and testing recognisers. This is the socalled *system-in-the-loop* method, in which users interact with an existing dialogue system while the data generated is collected. Obviously, there are limitations with the approach. First, it presupposes that a dialogue system is currently available to use for data collection. Second, it restricts the exercise to collecting data on usage patterns of a *current* system, when the planned *future* system may embody much increased functionality, or operate in a quite different domain.

Recommendation

System-in-the-loop data collections are useful for collecting speech data, and they may supply some baseline facts about how people use spoken language dialogue systems. However, this method should only be used for collecting more detailed data to guide future system specification if the functionality of the future system is planned to be only a small step beyond the current system which forms the basis of the data collection.

13.3.3 Design by simulation

Although it is certainly better to rely on analyses of human-human interactions than to rely on intuitions alone, for all but highly constrained menu dialogue systems, the fact remains that human-human interactions are not the same as human-computer interactions and it would be surprising if they followed precisely the same rules. The designer is caught in a vicious circle: it is necessary to know the characteristics of a dialogue between a person and an automaton in order to be able to build a system which acts as a dialogue participant, but it is impossible to know what such a dialogue would be like until such a system has been built.¹

This section examines how one particular simulation technique which has come to be widely used by dialogue system designers, the "Wizard of Oz" technique (see also Chapters 4, 9) can melp to extend designers' understanding of what human-computer spoken language dialogues would look like if only the systems which are currently at the planning stage were, in fact, implemented and running.

13.3.3.1 The Wizard of Oz technique

The basic idea behind the Wizard of Oz (WOZ) technique is simple: a human (usually known as the *wizard* or *accomplice*) plays the rôle of the

 $^{^{1}}$ We will encounter a similar problem when we come to consider system evaluation. It is possible to evaluate the users' perceptions of the system's usability but it is not possible to test its performance against a clear specification of what the system *should* be capable of, because this is, strictly speaking, unknown.

computer in a simulated human-computer interaction. It is not known who first coined the term in this context, though its etymology is obvious. In the children's novel *The Wizard of Oz* (cf. Baum 1900), the "great and terrible" Wizard turns out to be no more than a mechanical tin device operated by a man hiding behind a screen. As a corpus collection method, it is also less widely known as the PNAMBIC ("Pay No Attention to the Man BehInd the Curtain", from the film version of the *Wizard of Oz*) technique.

What primarily interests us here is the simulation of a computer system which takes spoken natural language input, processes it in some principled way, and generates spoken natural language responses. Example applications are telephone timetable inquiry services, hotel room booking services, home banking services and "intelligent" telephone answering machines. A survey paper published in 1991 (cf. Fraser and Gilbert 1991a) was able to assert that "very few WOZ experiments have attempted to simulate all the components of a speech dialogue system". However, since then there has been a dramatic increase in the number of groups using the technique to help specify interactive spoken language systems.

13.3.3.2 Requirements for WOZ simulations

WOZ simulations are only useful if certain conditions are met. The first condition is that the computer system being simulated is capable of being imitated realistically, given human limitations. For example, if it is known that the future computer system will need to undertake substantial database manipulation as part of its function, there is little point in setting up an unconstrained WOZ simulation, since people are not capable of performing such work within a realistic time period.

A second, less obvious precondition is that before the experiments are begun it should be possible to formulate a detailed specification of how the future system is expected to behave. This is necessary in order to ensure that the wizard is correctly simulating the intended system. This specification often needs to be more precise and more detailed than would normally be necessary just to build the computer system. For example, in a speech simulation, the wizard ideally needs to make recognition errors at the same rate and in the same way as the future system. However, while descriptions of speech understanding systems often specify error rates, they rarely indicate what kinds of errors are made, or in sufficient detail for the errors to be simulated. Indeed, one of the aims of using the WOZ technique may be to help devise such a specification. The way round this apparent paradox, that the design of the simulation requires a specification but the content of the specification depends on the results of the simulation, will be discussed later when we consider WOZ methodology.

A third condition for the usefulness of the WOZ methodology is that the task must ensure that the illusion that the wizard is a computer can be

convincingly maintained. In systems which communicate using text on terminals, only minimal precautions have to be taken, since the only evidence of the "computer" the subject sees is the output of characters on a screen (but even here, there may be value in buffering the output so that it appears a line at a time, rather than at the speed of the wizard's typing). In speech output channels, it is necessary to ensure that the wizard's speech is disguised to sound not quite natural, a condition often satisfied by use of a synthesizer. Similar problems arise in controlling the content of the wizard's output, which must use only knowledge likely to be available to a computer. The degree of attention which has to be paid to these issues is related to the likely gullibility (that is, likelihood of believing that the simulated system is real) of the subjects.

13.3.3.3 Variables in spoken WOZ experiments

In this section we consider some variables in spoken WOZ simulations. By "variables" we simply mean things which may vary. We make no distinction here between *control variables* which are set by the experimenter, response variables which are measured by the experimenter, and confounding factors, in which the experimenter has no interest or over which he has no control. The experimenter must decide how to treat variables in each simulation since there is considerable scope for variation between experiments. For example, in simulations of a telephone train timetable enquiry service, the caller's level of familiarity with telephone information services might be a confounding factor, producing significant differences between speakers. However, in an experiment which divides users into "experienced" and "novice" classes, this would be a control variable rather than a confounding factor. We shall restrict our discussion here to a straightforward listing of some of the variables in spoken WOZ simulations. For the purposes of our presentation, the variables can be divided into those relating to the subject, those relating to the wizard, and those relating to the communication channel.

13.3.3.3.1 Subject variables

Variables which concern the subjects in WOZ simulations can be subclassified into subject recognition variables, subject production variables, and subject knowledge variables.

SUBJECT RECOGNITION VARIABLES relate to the subject's ability to recognise the wizard's words.

• ACOUSTIC RECOGNITION Is the acoustic signal intelligible to the subject? The quality of canned and synthesised speech currently available ranges from fairly good to virtually unintelligible (see Chapter 12). The wizard's
speech should therefore display characteristics which locate it either somewhere on this spectrum, or just beyond the best available technology if the system being simulated is expected to include synthesisers currently at the design or development stages. The ability to understand synthetic speech is not constant; rather, it displays learning effects. Thus the ability to decode the acoustic signal is a variable, not just among speakers, but for a given speaker over time.

• LEXICAL RECOGNITION Does the subject recognise the words used by the wizard? The important question here relates not to acoustic recognition but rather to whether or not all of the lexical items used by the wizard are known to the subject. This variable could be expected to interact with the subject's domain expertise variable. For example, in a flight reservation application, the wizard might refer to an *apex* fare. If this word is not in the subject's vocabulary then he may not even know how to segment it (*apex, a pex, ape eggs ...*). The subject will either initiate some sort of breakdown recovery or he will adopt a wait-and-see strategy. The subject's unfamiliarity with items of the wizard's vocabulary is likely, sooner or later, to lead to clarification subdialogues which would not otherwise be present.

SUBJECT PRODUCTION VARIABLES relate to the speech and language produced by the subject insofar as they have implications for the ability of the wizard to recognise and understand the subject's words.

- ACCENT. A commercial telephone information service can not screen callers before they make their calls. A strong non-standard accent would cause problems for most currently available speech recognisers (assuming they have been designed or trained for a spectrum of accents centered around a perceived standard). If the wizard is to simulate a plausible future system then he must fail to decode strong accents in some principled way.
- VOICE QUALITY. Similar variability can be found in voice quality, but this time the variability is between individuals rather than speech communities.
- DIALECT. The subjects may manifest different dialects. Non-standard dialect words and more problematically non-standard syntactic forms would probably be unintelligible to the sort of computer system which can currently be envisaged.
- VERBOSITY AND POLITENESS. How direct are the subject's requests? What part does politeness play in the subject's talk?

SUBJECT KNOWLEDGE VARIABLES are concerned with what the subject knows.

• DOMAIN EXPERTISE. Concerning the application domain, subjects may have expertise which ranges from novice through to expert. The way in which the subject interacts with the system, the questions he asks of it, and the way in which he expects to be addressed by it, are likely to be affected by his level of domain expertise.

- SYSTEM EXPERTISE. Richards and Underwood (1984a) found that as subjects gained expertise in using a WOZ system, so they learned to frame requests more concisely and simply. Thus, the amount of system expertise a subject possesses is a significant variable.
- INFORMATION ABOUT THE WIZARD. What the subject is told about the wizard has an effect on dialogue structure and on the subject's view of the experiment. There is a body of evidence to show that people use different dialogue strategies according to whether they believe they are talking to a human or a machine (cf. Hauptmann and Rudnicky 1988). Speech to a computer has been labelled "formal" (cf. Grosz 1977), "baby talk" (cf. Guindon et al. 1986), "telegraphic" (cf. Guindon et al. 1987), and "computerese" (cf. Reilly 1987).

It seems unnecessarily complex to ask the subject to guess whether or not he is talking to a computer; this is to turn a simple WOZ experiment into a Turing test (cf. Turing 1950). The experiment would no longer be a simple simulation of future technology if the subject were given this additional discrimination task.

Thus, it seems that for routine simulations the subject should be led to believe that he is actually using the future technology. This can be expected to yield the best guide to how that technology will be used when it becomes available. Potentially there are ethical problems here since a responsible experimenter would not choose to tell an outright lie to the subject. A more appropriate approach is to tell the subject that the research aims to establish how people converse with computers, and to allow her/him to draw her/his own conclusions.

GULLIBILITY. What the subject is told is one thing, what he believes is quite another. In an experiment to determine the effect of awareness of the human operator on subjects' performance, (Newell 1989, p. 146) observes that:

It was pleasing to note that the subjects in the covert [i.e. misinformed] group all expressed surprise on being told that the experiment was based on a simulation. (Indeed, one [male] subject was substantially embarrassed on finding that a female operator had encoded the profanities which he had used when he was having difficulties and which had been faithfully reproduced on the screen!)

It seems that many subjects can be totally misled. Follow-up questioning can be used to determine what subjects believe about simulations. If they are not convinced then the results can be discarded.

It is interesting to note in passing one result of Newell's which appears to demonstrate the opposite of what might be expected (Newell 1989, p. 8):

Those subjects who were made aware of the operator's existence were more impressed... than those who thought they were talking to a computer.

13.3.3.3.2 Wizard variables

Wizard variables can also be divided into wizard recognition and production variables but these must be supplemented with extra classes of dialogue model variables and staging variables.

WIZARD RECOGNITION VARIABLES Corresponding to the subject's production variables are a set of wizard recognition variables defining the ranges of acoustic, lexical, syntactic and pragmatic phenomena which the wizard is allowed to recognise. One of the hardest tasks for the wizard is restricting what is recognised to what is defined by these variables. We shall see below (Section 3) one possible approach to formalising recognition constraints but, for the most part, the constraints will have to be applied directly by a wizard who knows his rôle intimately.

A tolerable error margin (i.e. of successful recognitions which should have been unsuccessful) should be set and any dialogues which, on postsimulation inspection, are found to stray beyond that margin should be discarded.

A particularly difficult problem is that of trying to mimic a speech recogniser which only manages to recognise the words in its limited vocabulary, and these only with, say, a 95% recognition rate. In order to be faithful to the technology the wizard would have to introduce a random (or partially random) 5% failure rate even with words which the system is supposed to know about. This is an almost impossible task. The best that can be expected is for the wizard to introduce occasional deliberate recognition errors. Of course, if the wizard is able to type the subject's words fast enough, an automatic system can be used to generate the appropriate errors with the target frequency.

The acoustic front-ends of speech-based information systems designed for use by the general public are likely to include rapid speaker adaptation capabilities. This means that speech recognition rates are likely to improve during the course of individual conversations. It is hard enough for a wizard to generate a fixed percentage of recognition errors; it would be virtually impossible for him to simulate an error rate which varies over time.

WIZARD PRODUCTION VARIABLES. Just like the subject, the wizard has production variables, but with the wizard these are defined by the performance of the existing or projected technologies.

Thus the whole gamut of speech generation variables (voice quality, intonation, syntax, register, etc.) need to be considered. Again, the wizard may be required to introduce principled errors at any of these levels if the simulation is to be faithful to the technology.

• RESPONSE TIME. One production variable of particular interest is the

wizard's response time. The object of a WOZ simulation should be to respond in more or less the same time as it would take the future system to respond and not in the time it would take a human to respond. Obviously, systems are planned to run in real time but the real time course of a human-computer dialogue is not yet known. It may be appropriate to allow a wizard to take slightly longer to respond than a human expert. The wizard will in any case require all the time he can get to apply conscious constraints to his normal recognition and generation capabilities.

In his listening typewriter WOZ experiments, Newell considered the question of response time to be so important that he trained his wizard to use a palantype keyboard (cf. Newell 1978) (an electronic stenography system which generates normal text) for rapid speech transcription (180 words/minute or more). The reasons why response times are important are simple (see also Chapters 8, 12 for further production variables):

- 1. Speed of response can be expected to affect dialogue structure and content.
- 2. Speed of response may also affect the subject's judgements of whether he is talking to a computer or to a human.

DIALOGUE MODEL VARIABLES. The model of the dialogue employed by the wizard is central to his interpretation of utterances and selection of responses to them. It is worth flagging the dangers of constructing a prototype dialogue model in advance of running simulations. The two-stage experiment carried out by Guyomard and Siroux (1987) indicates the amount of work required to define a minimally acceptable dialogue manager. In spite of their positive reports, it is to be expected that many simulation-analysisredesign iterations would be necessary to define a truly impressive dialogue manager. Since most research projects run to a tight schedule, a two-stage simulation is probably the best that most experimenters can hope for. STAGING VARIABLES. In this section we consider some practical matters

relating to the preparation of the wizard and the tools available to assist him in his work.

- TRAINING. The wizard requires training in at least three areas: the application domain, the system capabilities being modelled, and the tools available to assist in playing his role. The wizard should receive as much training as time allows in order to ensure that his performance is as close as possible to the projected performance of the future system.
- TOOLS. The wizard needs a lot of information at his fingertips. A range of tools could be designed to present this information as quickly and easily as possible. For example, a range of paper tools (charts, card indexes, etc.) and electronic tools (mouse menu systems, hypertext, etc.) could be used. A wizard's assistant might even be considered necessary.
- WIZARD PERSONALITY. So far we have assumed that the wizard is a person. But what if the wizard is part human, part machine? We shall call such a wizard a *bionic wizard*.

Kelley (1983a,b, 1984) proposes an iterative development scheme which involves running an initial WOZ simulation and then, in subsequent simulations, incorporating more and more subcomponents of the real system, moving in the direction of a more complex system in the loop. In the development of a speech input/output system this could involve placing a speech recogniser between the subject and the human wizard. Alternatively (or additionally), the wizard could respond with synthesised speech generated from text which could either be typed rapidly (e.g. on a palantype system) or selected from a file of standard responses. In principle, a bionic wizard could include many subcomponents of the system, with the human accomplice merely "plugging the gaps". A bionic wizard could expect to encounter a number of problems, not least of which is the lengthening of response time which a mixing of human and computer components might entail. However, if these difficulties can be overcome, iterative development represents a promising technique.

13.3.3.3.3 Communication channel variables

The simplest means of connecting the subject and the wizard is by telephone or similar two-way electronic communication channel. The quality of the channel can be yet another variable.

THE SUBJECT \rightarrow WIZARD CHANNEL. One way of modelling the performance of a speech recogniser is to degrade the subject's speech signal. This would save the wizard from the (almost impossible) task of consciously introducing recognition errors. However, the drawback of this method is that the wizard, who already has enough to cope with, is faced with the extra workload of interpreting degraded speech. The alternative presented above is to place a real speech recogniser between the subject and the wizard. However, if the object of the exercise is to simulate a future system, the use of existing technology might place unrealistic constraints on the simulation dialogues.

THE WIZARD \rightarrow SUBJECT CHANNEL. No subject is going to believe that they are talking to a machine if they are unable to distinguish its performance from that of a human speaker. An important part of the simulation is the "de-humanising" of the wizard's voice. One way to do this is to pass the signal through a vocoder to strip it of human intonation and make it sound "mechanical". A secondary effect might be to make it roughly as difficult for the subject to understand the wizard as it would be to understand a speech synthesiser. This similarity could never be better than approximate. The alternative to degrading the wizard's voice is to place a speech synthesiser between the wizard and the subject. Once again, the usefulness of this strategy depends, in part, on the extent to which the synthesiser approximates to the synthesiser in the projected future system. CHANNEL INTERACTION. Can signals pass in opposite directions at the same time? The reason why this is important is that it may be desirable to let either subject or system talk in overlap or interrupt the other. On the other hand, it may be desirable or necessary to prevent them from doing so. In either case, it is important that the capabilities planned for the future system should be designed into the WOZ simulation to ensure that turn-taking phenomena recorded in the experiments are relevant for the future system.

13.3.3.4 An iterative WOZ methodology

In this section we describe a methodology for using WOZ simulations to specify the functionality of a speech input/output system. The suggestions presented here draw heavily on the work of Kelley (1983a,b, 1984), Guyomard and Siroux (1986a,b, 1987, 1988), the SUNDIAL Project (cf. Peckham 1993) and the Danish National Project on Spoken Language Dialogue Systems (cf. Dalsgaard and Baekgaard 1994).

The methodology involves at least three phases: a pre-experimental phase, a first phase, and a subsequent phase or phases. The need for at least three phases in the methodology stems from the difficulty noted earlier, namely that a WOZ simulation is intended to simulate as exactly as possible a future computer system, but the requirements to be satisfied by that system (and thus its precise specification) may be one of the outputs of the simulation work. To get round this circularity, we propose an iterative methodology which over the course of several phases refines both the simulation and the system specification until, ultimately, they converge.

To begin, the simulation incorporates only gross features of the intended system, the wizard in other respects acting "normally", that is with full human capabilities. The first phase yields data which can be used to develop an initial specification of requirements and thus some constraints on the wizard's behaviour in the second phase. In principle, the cycle of simulation and specification could be repeated many times, but in practice, two or three phases are likely to be sufficient.

13.3.3.4.1 The pre-experimental phase

Before the simulation is carried out it is necessary to analyse the application domain in order to define the wizard's domain knowledge. This domain knowledge may be available on-line in the form of a database (e.g. a travel booking database). In this case the wizard must be trained to use the database query language.

A second pre-experimental task is to decide what the subjects are to be told and how they can be made to interact meaningfully with the system, without simply following a script. This problem can be overcome by the

use of *scenarios*: the subject is assigned a rôle and given some background information. The subject is then given a high-level description of a task to be accomplished (e.g. "you want to meet Aunt Matilda who is flying into London from Hong Kong this evening"). The subject is free to decide what needs to be asked (e.g. when flights are due, which airport she is arriving at, which terminal she is arriving at, etc), the order in which the questions should be asked, and the exact wording of the questions.

Thus a vital task in the pre-experimental phase is to design realistic scenarios which constrain the subject as tightly as possible to the application domain of the future system, while giving her/him as much motivation and as much freedom of expression as possible within these bounds.

At the pre-experimental phase a number of other practical matters need to be sorted out, such as:

- selecting a location for the experiments;
- installing the required hardware and software;
- finding subjects.

13.3.3.4.2 The first experimental phase

In the first experimental phase very few – perhaps no – constraints should be placed on either subject or wizard. Any constraints which are applied are likely to relate to what the wizard is allowed to say. The wizard's voice should, of course, be distorted so the interaction should not have the character of a free human-human conversation. This first phase should be used to gather data which can then be used in the definition of an initial lexicon, grammar, and dialogue model.

13.3.3.4.3 Second or subsequent experimental phases

The findings of the first phase should be used to define some constraints for the second phase. A clearer definition should now be available of what the wizard is not allowed to understand and what he is not allowed to say. In an ideal world it is conceivable that there could be many subsequent phases in which the insights of the last phase would be used to refine the current phase. It is also conceivable that at each iteration, a new or improved hardware or software component could be added to a bionic wizard, thus bringing the simulation ever closer, in fact as well as in appearance, to the future system.

13.3.3.5 WOZ conclusions

In summary, we have introduced the WOZ technique as a means of predicting the functional requirements of future spoken language dialogue systems. Though there are significant technical problems in setting up spoken WOZ simulations, with careful design a wizard *can* simulate a computer sufficiently well to fool almost all subjects almost all of the time. The fact that, with support, people can simulate future speech systems enables designs to be developed iteratively and evaluation to be carried out before significant resources have been invested in system building. This strategy of *early evaluation*, which has been recommended in other areas of computer system design, has obvious advantages of cost and speed of convergence to a satisfactory design over the only alternative: build, evaluate and re-build. We have identified a number of subject, wizard, and communication channel variables for spoken WOZ simulations. Taken together, these should provide an initial framework for staging and for comparing WOZ simulations.

13.3.4 Iterative design methodology for spoken language dialogue systems

Most of the spoken language dialogue systems which have been created so far (SUNDIAL, VODIS, PAROLE, etc.) have used analysis of real dialogues and simulated dialogues before proceeding to implement a system. These data have, of course, been augmented by designers' intuitions to fill genuine gaps in the data. For example, observation and simulation corpora in the travel information domain might not include mention of all the destinations contained in the timetable. The design should not be so tied to the data that these deficits can not be simply rectified. However, caution should be exercised in the use of intuitions so as not to equip the system with functionality which it will never need. Experience has shown that the expectation that some linguistic form *might* occur is not in itself sufficient grounds for supposing that it *will* occur.

Normal practice is to design several sequential versions of the system, each version benefitting from technology improvements and from analysis of results of earlier stages.

13.3.4.1 Interactive voice response systems: recommendations

13.3.4.1.1 Recommendations on design methodology

Designing a simple system-led menu-style small vocabulary interactive voice response system consists of the following steps, taking both human linguistic behaviour and speech technology performance into account.

- 1. Study the application domain and define what the tasks to be achieved are and what steps they consist of.
- 2. Translate the sequence of subtasks into a sequence of questions to be asked by the system and answered by the user, interleaved where necessary with system internal operations such as database lookup.
- 3. Define the exact wording of the system prompts, and the exact vocabularies and language models which are appropriate for each recognition.

- 592 Part III: Spoken language system assessment
- 4. Draw up a full specification of the IVR system, integrating the dialogue flow, system-internal operations, prompting and recognition constraints.
- 5. Design a first version (X) of the dialogue system.
- 6. Conduct laboratory tests with available technology using test corpora where available, and also laboratory staff simulating users.
- 7. Conduct field trials with real users, recording new corpora where deemed useful.
- 8. "Tune" the system by iteratively modifying, then testing it.
- 9. If too many modifications are necessary, respecify and reimplement the system.
- 10. Design an X+1 version of the system, integrating new technologies.
- 11. Carry out new laboratory tests with the new version.
- 12. Carry out field trials with real users.
- 13. Return to step 9 unless the system is deemed to be complete.

13.3.4.1.2 Recommendations on prompt design

Prompt design is especially important for IVR systems. Since the user has to follow the system's lead, that lead must be clear, unambiguous, and reassuring. The following recommendations summarise some simple steps which can be taken to achieve an effective prompting regime.

- 1. Keep prompts as brief as possible without being terse.
- 2. Keep prompts as simple as possible.
- 3. Use a consistent linguistic style for prompts.
- 4. Ensure that each prompt (except the last) finishes with an explicit question or command.
- 5. Wherever technically possible, allow users to interrupt the prompt.
- 6. Where prompt interruption is not possible, ensure that either the recogniser starts listening the instant the prompt stops playing, or use some audible signal to indicate when speech may begin.
- 7. If prompts are canned, either use a single speaker or, if more than one is used, ensure that each speaker serves an intuitively distinct function.
- 8. Do not expect instructions presented to the user at the start of a dialogue to be remembered in subsequent turns.
- 9. Wherever possible, re-promptings after errors or absence of user input should provide extra guidance to help the user behave in the desired fashion.
- 10. Control variables such as prompt voice quality to give the system a warm and friendly system "personality".

13.3.4.2 Spoken language dialogue systems: recommendations

13.3.4.2.1 Recommendations on design methodology

Designing a spoken language dialogue system consists of the following steps, taking both human linguistic behaviour and speech technology performance into account.

- 1. Study human-human interaction recordings in a situation similar to the one in which the system will be used, and make an ergonomic analysis of the needs or requirements of potential users.
- 2. Carefully define a Wizard-of-Oz simulation, making objectives explicit.
- 3. Conduct Wizard-of-Oz simulations (preferably using an iterative WOZ methodology) and record the complete resulting dialogues.
- 4. Transcribe the dialogues recorded in simulations, (several levels of transcriptions may be necessary). If possible use a standard transcription scheme.
- 5. Draw up a specification of the interactive dialogue system.
- 6. Design and implement a first version (X) of the dialogue system.
- 7. Conduct laboratory tests with available technology using corpora recorded in Wizard-of-Oz simulations, and then with laboratory staff simulating users, recording new data.
- 8. Conduct field tests with real users, recording new corpora.
- 9. "Tune" the system by iteratively modifying, then testing it.
- 10. If too many modifications are necessary, carry out new (bionic or human) Wizard-of-Oz experiments, allowing for controlling of different parameters.
- 11. Design and implement an X+1 version of the system, integrating new technologies.
- 12. Carry out new laboratory tests with the new version.
- 13. Carry out field tests with real users.
- 14. Return to step 9 unless the system is deemed to be complete.

13.3.4.2.2 Additional recommendations

In addition to these methodological guidelines, the spoken language dialogue specification/design process can be expected to be simplified and improved if a few extra recommendations are adhered to. (Many of these summarise points already made in the preceding discussion.)

- 1. Where time and other resources allow, base the specification on data from a diversity of sources.
- 2. Consult human-human data to learn about the task and to understand the service expectations which users will bring to the system.
- 3. Conduct WOZ simulations to determine the effect of human-computer factors for a specific task or application domain.
- 4. Use native speaker intuitions to fill obvious gaps in the human-human and WOZ corpora, but avoid going beyond this.

- 5. Use an iterative refinement methodology to perfect the specification.
- 6. Allow sufficient time and resources for the specification process.
- 7. Decide in advance which questions to ask of the data, and wherever possible stick to them.
- 8. Conduct a dialogue act analysis of the dialogues collected in the corpora, paying special attention to the conditions which must be satisfied in order to proceed from one dialogue state to the next.
- 9. Describe the dialogue state transitions using some formally explicit apparatus (such as a flowchart or formal specification language).
- 10. Use the data to identify the total lexicon required, then divide it into sublexicons, where each sublexicon is accociated with a dialogue act.
- 11. Use the data to identify a covering grammar, then divide it into subgrammars, where each subgrammar is accociated with a dialogue act.

Human reactions to spoken language dialogue systems have to be observed on the spot. The ideal approach is therefore to design systems in close collaboration with professional organisations which have groups of potential users who are willing to critique specification documents, participate in early trials, and feed back useful comments.

13.3.5 Readings in interactive dialogue system specification

Very few – if any – publications have presented thorough accounts of practical experiences of the entire process of spoken language dialogue system specification. A reasonable number of publications report laboratory experiments to collect data which might be useful as a basis for design. The following are some of the best places to look first for further information: Andernach et al. (1993), Andry et al. (1990), Dahlbäck and Jönsson (1986, 1989), Dalsgaard and Baekgaard (1994), Delomier et al. (1989), Diaper (1986), Diaper (1989), Dybkjaer et al. (1993), Fraser (1991), Fraser and Gilbert (1991b,a), Fraser et al. (1992), Guindon et al. (1986), Guindon et al. (1987), Guindon (1988), Guyomard and Siroux (1987), Guyomard and Siroux (1988), Hansen et al. (1992), Hauptmann and Rudnicky (1988), Jönsson and Dalbäck (1988), Kelley (1984), Labrador and Dinesh (1984), Leggett and Williams (1984), Life et al. (1988), Luzzati and Neel (1989), MacDermid (1993), Müller and Runge (1993), Olsen and Olsen (1990), Ponamale et al. (1990), Richards and Underwood (1984a,b), Tubach and Doignon (1991), Whittaker and Stenton (1989), and Wooffitt and Fraser (1992).

13.4 Evaluation

Few complete interactive dialogue systems have been systematically evaluated. Recommendations will primarily be based on evaluation tests which have already been performed over the past few years, most of which concern rather simple tasks. The lack of systematic evaluation lies in the fact that few interactive spoken language dialogue systems have yet been in real use, but also in the fact that there does not exist any stable categorisation of the basic units used in dialogue systems which might constitute a reference, as opposed to other systems (for instance, in electronic dictionaries, basic units – lexical entries – are assigned grammatical categories and other characteristic features). For dialogue systems, there has been no such definitive categorisation of dialogue acts, nor, for that matter, has there been any definitive understanding of "dialogue grammar", though state machines for turn sequencing are widely used. Furthermore, a dialogue system encompasses several different levels, each of which consists of different components, making an overall evaluation more complex.

The set of recommendations contained in this chapter should therefore be seen as provisional, awaiting further refinement and extension as understanding in the area of interactive dialogue systems grows.

13.4.1 Background

The last few years have brought a new awareness of the importance of evaluating speech and language technology according to accepted standards. This allows progress to be monitored within a single project; it also facilitates meaningful comparison across projects. In Europe, much of the enabling work on speech has been carried out as part of the SAM project (cf. Fourcin et al. 1989). This has concentrated on developing standards for the storage, labelling, and basic speech processing of acoustic data. Similar work has been carried out as part of the U.S. DARPA program in Spoken Language. Text-based systems have also been considered under the DARPA Written Language program. As well as looking at low level issues, these programmes have gone on to consider higher levels of speech and language processing. For example, in text processing, progress has been achieved in evaluating the coverage of grammars (cf. Black et al. 1991) and on the evaluation of text understanding systems (cf. Sundheim 1991). Spoken language understanding has been evaluated by monitoring the ability of systems to generate appropriate database queries on the basis of spoken questions (for example, see Bates et al. 1990; Pallet et al. 1990; Price 1990).

At the root of almost all approaches to evaluation of speech and language technology lies the notion of a *reference response* or *reference answer*. The performance of the system is judged using the standard of the reference answer. Thus, a speech recogniser's performance is evaluated against what was actually said. A text understander is judged according to its ability to fill slots in a reference frame constructed on the basis of experts' judgements. A speech understander is judged according to its ability to construct for an utterance the same database query as a panel of experts. In each case, it is possible to prepare – in advance of any trials – a database of paired

tasks and reference answers. This greatly simplifies the task of objectively comparing different systems.

The "reference answer" approach does not extend straightforwardly to the evaluation of dialogue systems, whether they use spoken or written language. First, dialogues are *complex* structures: they may accomplish multiple tasks for which multiple metrics are required. Note, however, that a certain amount of misunderstanding is a normal feature of successful human dialogue. The success of parts of a dialogue is subordinate to the success of the dialogue as a whole. Second, dialogues are *dynamic* structures: the overall structure of the dialogue emerges out of the interaction of system and user, where each utterance is contingent upon those which precede it. This makes it very difficult indeed to construct meaningful reference material.

These problems have led to dialogue systems being evaluated in a relatively simplistic fashion. For example, the final evaluation of the VODIS voice operated database inquiry system looked at the percentage of completed tasks and the mean time for completion, the reasons why dialogues were abandoned, the number of words in subject utterances, the word recognition rates, and an analysis of those instances in the trial when the system recognised nothing (cf. Cookson 1988). In addition, a questionnaire was used to elicit subjects' perceptions about the usability of the system. While all of these results are interesting descriptions of aspects of the system, taken together they do not present a clear picture of the system's capabilities. This work is presented as an exemplar of a class of similar evaluations. There is no intention to single it out for special criticism.

One proposal for obtaining a measure of the effectiveness of a system quadialogue participant is to look at its ability to understand an utterance in a dialogue context. Hirschman et al. (1990) have proposed a methodology which involves setting up a database of paired tasks and reference answer, where a task consists of an utterance plus an encoding of a dialogue state. This allows the system to be reinitialised between turns in a dialogue. In this way, the problem of the dynamic nature of dialogue can be managed. Different dialogue systems can be compared, so long as they are presented with the same utterance + canonical context pair as input. A similar approach was developed in the SUNDIAL project for testing and debugging the system. An extension of this approach - the dialogue breadth test - tests a system with a broad range of different utterance types for each canonical context, thus exploring the ability of the system to cope with the lack of constraint on next user utterance which exists at many points in a dialogue (cf. Bates and Ayuso 1991). The canonical context approach represents a significant improvement on previous reference answer approaches. However, it only evaluates the ability of a system to perform context-sensitive interpretation; it focusses on the abilities of a system to interpret local structures in dialogue. But dialogue consists of larger structures, and these are beyond the scope of this metric. A dialogue system could perform reasonably well on this metric but be incapable of completing a single dialogue successfully. One approach to evaluating the abilities of a dialogue system to deal with the larger structures of interaction might be to monitor the abilities of such a system to recreate an entire reference dialogue. For example, a corpus of dialogues could be collected using the Wizard of Oz simulation methodology discussed above. Subjects could then be asked to accomplish the same tasks using a dialogue system, and the results of the two exercises could be compared. Bates and Ayuso have argued convincingly that such an approach is unrealistic; in fact, they go so far as to compare it to "asking one chess expert to exactly reproduce every move that some other expert made in a past game!" (Bates and Ayuso 1991, p. 320). While accepting this general point, Fraser (1991) has claimed that there is some merit in comparing the results of simulations and system data collections in a more sophisticated way. This involves analysing the simulation corpus and generating an abstract multilevel description of it. This has the effect of defining at a theoretical level the space within which reasonable system behaviour may be located. A similar analysis of the system corpus is carried out and the results are compared. In the language of Bates and Ayuso's chess analogy, this is like comparing two games of chess and observing that, though they may differ in detail, they both include a version of the Sicilian Defence. However, to the best of our knowledge, this approach has not yet been thoroughly tested.

13.4.2 Characterisation

Characterisation is a very important part of the system evaluation process. It is vital to define precisely what it is that is being evaluated, together with all the conditions under which the evaluation takes place. We therefore devote attention here to characterising the dialogue system, the task, the user, the environment, the corpus, and the overall system.

13.4.2.1 Dialogue system characterisation

Dialogue systems may be characterised by several parameters, defining their complexity; these are listed in appropriate categories below.

13.4.2.1.1 Knowledge databases

LANGUAGE MODEL: the model used by the recogniser and shared by the system in order to guide the recognition process, whenever necessary.²

 $^{^{2}}$ This use of the term *language model* should not be confused with current usage in speech technology to refer to stochastic grammars or word context models which limit the search space for word hypotheses in automatic speech recognition. See Chapter 7. [Technical Editor]

- 598 Part III: Spoken language system assessment
- RECOGNISER COMPLEXITY: isolated word, word spotting, continuous speech (error-free read text or spontaneous with repairs, hesitations, ill-formed or incomplete sentences).

LEXICON: the list of allowed words.

PHONOLOGICAL RULES: rules of pronunciation of the words.

- SYNTACTIC RULES: descriptions of the well-formed linguistic constructions recognised by the system.
- SEMANTIC-PRAGMATIC REPRESENTATION: the list of concepts used with associated structures (frames, conceptual graphs, etc.).
- TASK MODEL: plans and scenarios, goals and subgoals, representation of the objects of the task and of their characteristics. (For example, in an air-traffic control domain, known objects would include planes with parameters such as heading, level, etc., and known goals would include authorising route adjustments, etc.).
- DIALOGUE GRAMMAR: hierarchy of subdialogues with dialogue act leaf nodes, indicating permitted progression from one turn type to the next.
- USER MODEL: particular rules of pronunciation (confusion matrix), linguistic behaviour (particular formulations), user beliefs and knowledge about the task, etc.
- SYSTEM MODEL: the list of the media available, with a description of their characteristics.

13.4.2.1.2 Dialogue strategies

The following are alternative classes of dialogue strategy which may be adopted. Considerable scope exists for further subclassification here.

- STRICTLY GUIDED AND DETERMINISTIC: no initiative left to the user. IVR systems typically fall into this category.
- COOPERATIVE: includes correction and prediction mechanisms, shares initiative with the user, accepts interruptions or negotiation, capable of clarifying the system's choices and responses, (turn-taking is balanced between the user and the system)
- CONSTITUTIVE: (for educational systems) the system has to learn new notions in its normal operation.
- ADAPTIVE: takes into account the dynamic user model by learning the users communicative strategies and adjusting to them as each dialogue proceeds.

13.4.2.2 Task characterisation

There is an intimate connection between the application domain and associated tasks, the language required to accomplish these tasks in this domain, and the design of a system which supports dialogues for this purpose.

13.4.2.2.1 Task type

A broad categorisation of task types can be made, depending on whether the objects of the task are evolving during the dialogue or not. These include the following:

- INFORMATION ACCESS AND RETRIEVAL: for example, train or flight time table enquiries.
- NEGOTIATION: the system acts as an expert-system, trying to find the best solution, for instance the best way to assign conference delegates to hotels, taking hotel costs and proximity to the conference centre into account. (An information retrieval system may need some kind of negotiation, for example, to obtain a less expensive travel ticket).
- $\ensuremath{\mathsf{Process}}$ control: the task is evolving, as for instance in communication with a robot.
- TRAINING: knowledge acquisition by the user or by the machine. In such cases as air-traffic control training, the task may be evolving (planes are changing heading or level).
- MONITORING: the system does not play an active part in the dialogue, but monitors its progress and is available to offer assistance when called upon. Such systems are sometimes referred to as computer mediated (or supported or assisted) human-human communication systems. The most notable example of such a system is the VERBMOBIL face-to-face spoken language translation system, currently under development.

13.4.2.2.2 Task complexity

Tasks and the dialogues by which these tasks are achieved are more or less mutually dependent. Typically, simple tasks will be solved by means of simple dialogues and complex tasks will be accomplished by means of complex dialogues. Thus an important part of the characterisation of the dialogue system is an index of the complexity of the task or tasks to be addressed by the system. Such indices might include the following subcomponents.

- the number of different scenarios covered (i.e. does the system address just one kind of problem or many different kinds?);
- the maximum complexity (i.e. in each scenario, what is the maximum number of subgoals which have to be satisfied in order to solve the task problem?). The complexity may be measured by the depth or width of the hierarchy, if the task and subtasks can be represented by a tree structure;
- number of subtasks to be achieved in parallel (especially in multimodal interaction);
- the minimum number of exchanges necessary to solve the task problem or complete a plan.

13.4.2.3 User characterisation

A large number of different criteria must be taken into account when characterising users of interactive dialogue systems. At least the following must be considered:

- number of users, for example, a few (10), or numerous (thousands);
- age (children, adults). The following age bands are adequate for most assessment purposes: less than 18, 18-25, 25-35, 35-45, 45-55, 55-65, over 65;
- sex (female/male);
- experience in the use of the automatic system (trained or untrained, experienced or novice, occasional or regular users);
- expertise in the application domain (the user knows what information he wants or not);
- status (professionals or members of the general public);
- motivation (whether they are real end-users, or paid or unpaid subjects);
- physical status (stressed, tired, ill, ...); when stressed due to adverse environments (for instance, in a space shuttle), the user might be affected by vibrations, temperature, G-effect, urgency, etc. in his pronunciation and utterance structure, and in his way of conducting the dialogue (he may wish to complete the task very quickly, for instance).

13.4.2.4 Environment characterisation

The environment is the total context in which an interactive dialogue system is evaluated. In general, it would be impossible to produce an exhaustive description of an environment, but a restricted set of relevant features may be selected usefully. For example, the following features relating to the acoustic environment might be used.

- type and proximity of the microphone used by the recogniser (for example, high quality microphone, close-talking microphone, microphone array, telephone handset, hands-free telephone, etc.);
- level of background noise (anechoic chamber/ office/ street/ car/ factory);
- communication quality (telephone lines: analog/digital).

Other relevant environment features will also have to be developed. As a general rule, the more features of the environment which can be characterised, the better.

13.4.2.5 Result corpus characterisation

A system evaluation will result in the collection of a corpus of resulting dialogues. Corpora must be fully characterised to ensure that changes in system performance over time can be tracked, and that corpora collected using different systems can be reliably compared.

At least the following features of an evaluation corpus should be noted:

- length of the corpus (in terms of elapsed time);
- number of different speakers;
- number of scenarios per user, number of identical scenarios processed by different users;
- length of each scenario or average length;
- number of dialogues, utterances, words, etc.;
- number of words per utterance (average), etc.;
- type of environment in which it has been recorded (and how far is it from the target usage conditions?).

13.4.2.6 Overall system characterisation

The overall system in which the dialogue system is embedded also needs to be characterised. First, though, it is necessary to clarify some terms.

A mode refers to perception senses which allow for communication: the following modes may be identified: vocal, visual, auditive, tactile, olfactive. *Communication means* (or *media*) refer to materials or devices which are used by the dialogue system to communicate with the user.

Communication modalities concern the way the communicating agent/party uses a mode: for speech, different modalities may be identified, for example whether continuous speech or isolated words are used, whether a whispering or shouting style is used, etc.

The system may comprise different communication means which may be characterised by:

- number of different media;
- media usage supported (in parallel, combined, alternate, etc.).

Each medium has an associated language model and is characterised by:

- medium information processing time;
- availability;
- input/output modalities: for example, for a recogniser, output to the system might be words or sentences; for a synthesiser, input from the system might be sequences of phonemes or conceptual graphs.

This characterisation is particularly important in multimodal dialogues, as the system's awareness of the states of each of the media (active, available, occupied, etc.), at each step of the dialogue, is determinant in the system predicting which media will be used by the user, or choosing the adequate media to send information to the user.

13.4.3 Assessment framework

It is unlikely that a single measure could be used meaningfully to sum up the quality of an interactive language system in the foreseeable future, due to the large variety of dialogue systems and the complexity of their

different components. It is more likely that systems will be characterised by vectors of metrics, each one picking out a different aspect of the system's performance. Some of these aspects can readily be assigned a numeric value, whereas others are clearly qualitative.

13.4.3.1 Assessment metrics

Different types of evaluation must be identified depending on:

- the environment in which tests take place;
- the degree of simulation or system integration;
- how much is being evaluated.

13.4.3.1.1 The environment in which tests take place

The broad category of environment in which tests take place (laboratory tests against field trials) is of vital importance, and the selection of this environment will depend on the purpose for which the evaluation results are required. There are competing imperatives here. On the one hand it is valuable to be able to repeat experiments exactly, changing only the desired variables. The way to achieve this is to carry out laboratory tests with pre-recorded databases. (This is easier said than done in interactive systems, since there may be many different routes through a dialogue and even modest changes in experimental variables may cause the pre-recorded utterances to be out of phase with the system's utterances). On the other hand, since users contribute about half of every dialogue, it is important to trial dialogue systems in the field with real users operating under target usage conditions.

13.4.3.1.2 The degree of simulation or system integration

The degree of simulation or system integration (pre-recorded databases, Wizard-of-Oz versus system integration tested with real users). WOZ simulations are frequently used to test dialogue system specifications in advance of implementation. Likewise, simulations in which some components are real and others are simulated are used to test system integration plans. Unless there are good grounds for doubting it, it is reasonable to suppose that the same evaluation standards should be usable when all, part, or none of a system is being simulated.

13.4.3.1.3 How much is being evaluated?

GLASS BOX (diagnostic) evaluation considers the performance of one or several subcomponents of a dialogue system. The objective is to evaluate subcomponents in the context of a whole dialogue system, and diagnose the contribution of each part to the overall success or failure of the system.

BLACK BOX (performance) evaluation considers the overall performance of a dialogue system without reference to any internal components or behaviours.

13.4.3.2 Glass box assessment

The objective of a glass box evaluation is to evaluate each component as it serves its function in the whole system. This involves determining the gross performance characteristics of the major subcomponents (such as the recogniser, parser, semantic analyser, dialogue manager, message generator, and speech synthesiser). Further information on the assessment of these core technologies can be found in Chapters 10, 11, 12.

However, it should, in principle, also be possible to monitor certain more fine-grained internal features of the system's performance which relate more directly to the systems rôle as an *interactive dialogue* system, rather than simply a spoken language processing system. For example, the following features could profitably be investigated:

- The system's ability to:
 - interrupt synthesis,
 - shift focus or restart,
 - internally correct sentences or detect incoherences,
 - anticipate next exchanges,
 - resolve references (ellipsis, anaphora) and ambiguities (assessment of the dialogue history),
 - accept different user's formulations for the same semantic meaning,
 - accept ill-formed sentences, self-corrections, incomplete sentences,
 - process unknown words,
 - infer new data or information,
 - handle several dialogues or tasks in parallel (several users, several communicating entities), and
 - deliver equivalent outputs using different output media, according to the context of interaction.
- accuracy/pertinence of each answer;
- average number of repetitions to obtain the correct response (or action) (assessment of the way dialogue handles correction in interaction with the user);
- number of recognition errors corrected by higher-levels of comprehension (assessment of system's internal predictive and corrective mechanisms).

This is currently little more than a "wish list" since very few results in any of these areas have yet been achieved or published. However, readers working on interactive dialogue systems are encouraged to consider these questions and contribute findings or lessons learned to help extend current levels of knowledge.

13.4.3.3 Black box assessment

In a black box assessment exercise, the interactive dialogue system is treated as an informationally encapsulated module. It is possible to monitor inputs and outputs, but not to look inside the box. Black box metrics are appropriate for characterising whole systems, and system comparisons should take place at the level of black box results.

No standards for black box assessment of interactive dialogue systems have yet emerged. However, the areas mentioned in the following sections should be considered as candidates for this kind of analysis.

13.4.3.3.1 Quantitative measures

- Average number of exchanges to obtain relevant responses or reactions from the system;
- task completion rate: the percentage of all tasks which the system succeeds in completing;
- transaction success rate: the percentage of all tasks which the system succeeds in completing with the correct answer;
- interaction speed (response time of the system);
- terseness of system's answers (average number of words per system utterance – whether a high or a low average is deemed to be desirable is likely to vary from application domain to application domain).

13.4.3.3.2 Qualitative measures

• User satisfaction: a number of projects have used questionnaires, with apparent success, to elicit subjective evaluations from users. Users are typically asked to evaluate the system for some subjective qualities (pleas-antness, usefulness, etc.) and to compare the system to more traditional means of communication, such as human-human interaction. A simple method is to present users with a five point scale, from which the best fitting choice should be selected. For example:

Compared to using an existing information service, using the computer information system is... [circle choice which is most appropriate]

Much		the		much
easier	easier	same	harder	harder

- ability to adapt to new users (this, and the following capability to be judged on external evidence only),
- ability to adapt to the same user,
- ability to handle multi-modality.

13.4.3.4 A core set of metrics for system comparison

It should be clear from the somewhat schematic discussion of interactive dialogue evaluation metrics that the field is still at a fairly primitive stage of development. Therefore, in this section we describe a core set of evaluation metrics (all of them black box metrics) which can be used, in the interim, in order to provide a comprehensible and concise characterisation of the system's capabilities. The proposed set is far from complete, and must be regarded as no more than provisional. However, if the metrics are taken up and applied to a number of different systems, it should be possible to learn a reasonable amount about the performance of a given system relative to other systems assessed (under near-identical conditions) using the same set of metrics.

The core metrics to be employed are these:

DIALOGUE DURATION: the average duration of a dialogue.

TURN DURATION: the average duration of a turn.

- CONTEXTUAL APPROPRIATENESS: a measure of the appropriateness of the system's turn-by-turn behaviour.
- CORRECTION RATE: a measure of the proportion of all turns which are devoted to rectifying problems.
- TRANSACTION SUCCESS RATE: a measure of the percentage of all transactions which the system completes appropriately.

These metrics are described in more detail below.

13.4.3.4.1 Dialogue duration

SHORT NAME: DD

DEFINITION

Dialogue duration is a measure of the average duration in seconds of a dialogue.

Methodology

Ensure that all dialogues in the evaluation corpus are timed. A good way to do this is to get the system to keep a record of its "connect time" (time when it is being used). To calculate DD, divide the total amount of dialogue connect time by the number of dialogues in the corpus.

13.4.3.4.2 Turn duration

SHORT NAME: TD

DEFINITION

Turn duration is a measure of the average duration of one turn in a corpus of dialogues.

Methodology

The methodology proposed here is for a minimal TD measure. To obtain a TD figure, divide the total amount of dialogue connect time by the total number of turns in the corpus (where a turn is a contiguous block of speech contributed to a dialogue by either the system or the user). This figure averages across system turns (measured from when the user stops speaking to when the system stops speaking) and user turns (measured from when the system stops speaking to when the user stops speaking). Some researchers may wish to distinguish between these. However, for the baseline set of metrics described here, it is proposed that the simpler calculation be used.

13.4.3.4.3 Contextual appropriateness

SHORT NAME: CA

DEFINITIONS

Contextual appropriateness is a measure of the appropriateness of a system utterance in its immediate dialogue context. This is a five-valued measure, with values drawn from the set:

• TF (total failure)

Total failure is recorded when the system fails to produce any linguistic response to a user utterance.

• AP (appropriate)

A complete definition of appropriateness is not practically possible. Roughly speaking, an utterance is deemed to be appropriate if it does not violate any of Grice's maxims for cooperative conversation (cf. Grice 1975), as re-interpreted in the context of human-computer dialogue (cf. also Gibbon 1992b).

Intuitively, an utterance is appropriate in context if it is not unexpectedly conspicuous (marked) in some way. An utterance can be marked in a number of ways such as by introducing an unnatural or non-sensical progression of dialogue acts, or by being uncooperative, or by being noticeably over- or under-informative. This is a first order metric which groups together a number of phenomena. Analysis of these results should lead to the development of some finer-grained second order metrics in the future as understanding of the key issues builds. Contextual appropriateness values are expressed as percentages of all system utterances.

- IA (inappropriate) An utterance is deemed to be inappropriate in context if it violates one of the conversational maxims.
- AI (appropriate/inappropriate) An utterance is classed as appropriate/inappropriate if the appointed experts cannot reach agreement on its contextual appropriateness.
- IC (incomprehensible) An utterance is categorised as incomprehensible if the experts are unable to make an appropriateness judgement because the content of the system utterance cannot be discerned. This may happen for a variety of reasons, e.g. an error in the text-to-speech system.

Methodology

It is difficult – perhaps too difficult – to specify *a priori* the range of all possible system utterances, in part because it is impossible to anticipate *a priori* the full range of user utterances for which a response will be required. Thus, contextual appropriateness scoring will be carried out by a "panel of experts". Two "experts" (e.g. members of a project team, though this may lead to overrating, cf. Chapter 9) will independently score each system utterance in a corpus. Where both experts agree, the scores will stand. Where the experts disagree, they will try in the first instance to reach agreement by discussion. Where this still fails to produce agreement, the utterance will be classified as AI (appropriate/inappropriate), thus flagging the fact that there is genuine uncertainty which requires further investigation.

EXAMPLES

(These illustrative examples are drawn from the travel information domain.)

EXAMPLE CA.1: TF [total failure]

U: What time does that leave? [closedown]

EXAMPLE CA.2: IA [inappropriate dialogue act]

U: What time does the next train leave for Milan?

S: Can I help you?

EXAMPLE CA.3: IA [uncooperative answer]

EXAMPLE CA.4: IA [under-informative answer]

- U: What time does the next train from Rome arrive?
- S: In the afternoon.

EXAMPLE CA.5: IA [over-informative answer]

- U: What time does the next train from Rome arrive?
- S: The next train from Rome leaves at 11 o'clock in the morning. It stops at Florence at 12.30. It arrives in Turin at 3 o'clock in the afternoon. The train consists of seven cars, drawn by an electric engine (serial number B475). Refreshments will be available, between 11.30 and 14.45. All major credit cards accepted.

EXAMPLE CA.6: IC [unintelligible answer]

- U: What time does the next train from Rome arrive?
- S: rhubarbrhubarbrhubarb

As pointed out above, CA is a first order metric. Though each of the examples classed as IA is inappropriate in some way, some seem more profoundly bad than others. So, for example, CA.2 is non-sensical, whereas CA.4 is just extremely curt. Two things are worth bearing in mind. First, CA is just one metric amongst several, and we can expect the categories used by other metrics to cut across IA. Second, notwithstanding the general issues relating to Grice's Co-operative Principle, judgements of contextual appropriateness must be earthed in a system specification. An unco-operative answer may not be disasterous for the flow of the dialogue but, given some specification of a cooperative spoken language dialogue system, and un-cooperative answer may be judged to be just as inappropriate as a non-sensical one.

13.4.3.4.4 Correction rate

SHORT NAME: CR

DEFINITION

The correction rate is the percentage of all turns in a dialogue which are concerned primarily with rectifying a "trouble".³ In general, turns which introduce troubles and those which correct them have the status of *insertion sequences* – they interrupt the flow of the dialogue without contributing new propositional content to it. (They may, of course, make substitutions in the propositional content.) If these sequences were removed, the dialogue would retain the same basic informational content and progression.

System turns which seek to correct a user misunderstanding about the capabilities of the system should not be included in the measure as a correction turn. The rationale is that when users try from the outset of a dialogue to misuse the system, dialogues can be very short with all of the system's

 $^{^3}$ "Trouble" is a general term to describe a broad class of dialogue problems such as those caused by speech recognition or parsing failures, misunderstandings, illogical or inconsistent utterances or belief states, etc.

utterances devoted to correcting the user's misapprehensions. This would skew the figures badly, though the user might have been acting reasonably in ignorance of the system's capabilities and the system might have coped perfectly.

The correction rate could be divided into two subsidiary measures: the "system correction rate (SCR)" and the "user correction rate (UCR)". Definitions of these rates are as follows:

CR: Percentage of all turns which are correction turns

 SCR : Percentage of all system turns which are correction turns

 $\mathrm{UCR}\colon$ Percentage of all user turns which are correction turns

Methodology

Once again, this metric can only be applied subjectively. Two experts independently count the number of system and user turns in a set of dialogues and keep a record of the system and user correction turns. Any conflict between experts will either be resolved by discussion or the offending turn will be removed from the measure and stored for further consideration. Once all of the uncertain turns have been removed, normalisation must take place in order to obtain a reliable CR rate. If n more system turns have been removed than user turns or *vice versa*, then the total number of turns must be reduced by n. This is not relevant in the case of SCR or UCR. Examples

(S = system correction turn; U = user correction turn.)

Example CR.1

- S: Flight information. Can I help you?
- U: What time does flight BA 123 arrive from Paris?
- S: BA 923?
- U: No. BA 123.
- S: BA 123?
- U: Yes.
- S: BA 123 from Paris leaves at 7.15am.
- \mathcal{U} : What time does it arrive?
- S: That flight arrives at London Heathrow at 7.20am.
- U: 7.28?
- \mathcal{S} : 20 past 7 in the morning.
- U: OK. Thanks.
- S: Thank you. Goodbye.

No. of turns $= 13$	No. of corrections	= 3
No. of system turns $= 7$	No. of \mathcal{S} turns	= 1
Number of user turn s 6	No. of \mathcal{U} turns	= 2

CR = 23%SCR = 14%UCR = 33%

Under normal circumstances the correction rate would be calculated for a corpus rather than for a single dialogue.

Example CR.2

- S: Flight information. Can I help you?
- U: I'd like to book a flight to Genoa.
- S: I'm sorry, this is just an information service. For reservations dial 071–340 4000.
- U: OK. Thanks.
- S: Goodbye.

None of these turns counts as a correction turn for the purposes of this metric. Thus the CR = 0%.

13.4.3.4.5 Transaction success

Short name: $\mathbf{T}\mathbf{S}$

DEFINITION

Transaction success is a measure of the success of the system in providing users with the information they require, if such information is available in the database. This is a four-valued measure:

S: succeed

- SC: succeed with constraint relaxation
- SN: succeed in spotting that no answer exists

F: fail

TASKS FOR WHICH ANSWERS EXIST. A "reference answer" can be associated with most task scenarios in advance of the dialogues based on the scenario. A reference answer consists of a frame with some or all slots filled in. Some of these are marked as obligatory – the transaction can only be judged to have succeeded (S) if the system conveys to the user the information stored in these slots. Other slots are marked as optional – the transaction will succeed even if the system does not tell the user the information in these slots. A transaction must be judged to have failed (F) if the system provides to the user any information which is inconsistent with that found in the reference answer frame, or if it fails to provide obligatory information to the user.

For example, here is a flight information scenario and associated answer frame. (Slots marked with an asterisk must be filled in a successful answer).

Assessment of interactive systems 611

Scenario 1	
Find out when flight BA 123 from Paris arrives.	

Reference Answer Frame	1
TASK:	flight enquiry
FLIGHT ID:	BA123
FROM CITY:	Paris
FROM AIRPORT:	Charles de Gaulle
TO CITY:	London
TO AIRPORT:	Heathrow
TO TERMINAL:	4
DEPART TIME:	15.35
ARRIVE TIME:	16.00*

Thus, any transaction in which the system tells the user the arrival time and does not contradict any of the other slot-fillers will succeed.

If a user introduces information which was not contained in the scenario and not anticipated in the reference answer frame, then an expert must produce a *post hoc* reference answer frame using this information and the task success must be judged in the light of this reference answer frame.

TASKS INVOLVING UNKNOWN OBJECTS WHICH CAN BE FOUND BY RE-LAXATION. If the user asks the system to perform a task type within its general competence, but the user references non-existent objects, then the transaction will be judged to have succeeded (SC) if the system is able to relax constraints until the user accepts an answer relating to the closest known object, and that information is correct according to the database.

TASKS INVOLVING UNKNOWN OBJECTS WHICH CANNOT BE FOUND BY RELAXATION. If the user asks the system to perform a task type within its general competence, but the user references non-existent objects which cannot be found by constraint relaxation, then the transaction will be judged to have succeeded if the system informs the user of the non-existence of the object and the user either ends the dialogue (SN) or replaces the non-existent object with one which exists (S) or can be relaxed satisfactorily (SC).

TASKS FOR WHICH NO ANSWER IS KNOWN. If the user asks the system to perform a task beyond the planned competence of the system then the reference answer frame will include the entry:

Reference Answer Frame	2
TASK:	unknown

The dialogue will be judged to have succeeded (SN) if the system informs the user that it is unable to perform the task requested.

In all other circumstances, the task will be judged to have failed (F).

Transaction success is a very simple measure of whether or not the system as a whole has come up with "the right answer".

Methodology

Where possible, an expert will prepare a reference answer frame for each scenario in advance of data collection. Where this is impossible (as in the circumstances described above), an expert must construct a reference answer frame on the basis of the task which the user appears to be trying to solve and on the basis of information in the database. The success of the task must then be judged against the new reference answer frame.

EXAMPLES

The following examples assume Scenario 1 and Reference Answer Frame 1, above.

EXAMPLE TS.1: S

- S: Flight information. Can I help you?
- U: Hello. Can you tell me what time BA 123 gets in from Paris this afternoon?
- S: BA 923?
- U: No. BA 123.
- S: BA 123?
- U: That's right.
- S: That flight leaves Paris at 3.35pm and arrives at London Heathrow airport at 4pm.
- U: Ok. Thanks a lot.
- S: Thank you. Goodbye.

EXAMPLE TS.2: SC

- S: Flight information. Can I help you?
- U: Hello. Can you tell me what time BA 923 gets in from Paris this afternoon?
- S: BA 923?
- U: That's right.
- S: There is no flight BA 923. Perhaps you mean flight BA 123 which leaves Paris at 3.35pm and arrives at London Heathrow airport at 4pm.
- $U{:}\quad Oh.\ OK.\ Thanks \ a \ lot.$
- S: Thank you. Goodbye.

EXAMPLE TS.3: SN

- S: Flight information. Can I help you?
- U: Hello. I'd like to book a flight on this afternoon's flight BA 123 from Paris to London.S: I'm sorry. This is just an information service.
- For reservations dial 071–340 4000.
- U: Oh. OK. Thanks a lot. S: Thank you. Goodbye.

EXAMPLE TS.4: F

- S: Flight information. Can I help you?
- U: Hello. Can you tell me what time BA 123 gets in from Paris this afternoon?
- S: BA 923?
- U: No. BA 123.
 - [closedown]

EXAMPLE TS.5: F

- S: Flight information. Can I help you?
- U: Hello. Can you tell me what time BA 123 gets in from Paris this afternoon?
- S: BA 123?
- U: That's right.
- S: That flight arrives in London at 5.30pm this evening.
- U: OK. Thanks very much.
- S: Goodbye.

13.4.4 Recommendations on evaluation methodology

In this section we summarise the evaluation methodology proposed here and present some concrete recommendations (most of them recapitulating points raised in the preceding discussion) for carrying out effective and reliable interactive dialogue system evaluations.

The evaluation methodology can be expressed by means of a series of recommendations. These should be read in sequence.

- 1. The sequence of events to be followed in evaluating an interactive dialogue system is: characterisation, data collection, analysis and application of metrics. Try to keep these as discrete, non-overlapping phases of work, as this helps to ensure that the test is as fair as possible.
- 2. Provide a characterisation of all relevant aspects of the dialogue system, the task, the user, the environment, the corpus, and the overall system. Most of this can be done before the data collection phase, though certain pieces of information (e.g. relating to users and corpus characteristics) will necessarily have to wait until the data collection is under way.
- 3. Produce a clear statement of the objectives of the evaluation exercise prior to the start of that exercise.
- 4. Select the minimum set of metrics which will satisfy the evaluation objectives.

NAME		VALUE	
DD	Dialogue duration	average dialogue duration in seconds	
TD	Turn duration	average turn duration in seconds	
CA	$\operatorname{Contextual}$		
	${ m Appropriateness}$	TF total failure	
		AP appropriate	
		IA inappropriate	
		AI appropriate/inappropriate	
		IC incomprehensible	
CR	Correction Rate	percentage of turns which are corrections	
SCR	System Correction Rate	percentage of system turns which	
		are corrections	
UCR	User Correction Rate	percentage of user turns which	
		are corrections	
TS	Transaction Success	S succeed	
		SC succeed with constraint re-	
		laxation	
		SN succeed in spotting that no	
		answer exists	
		F fail	

Table 13.1: Summary of core comparative evaluation metrics

- 5. When budgeting time and personnel for the evaluation task, be sure to plan adequate resources to complete the task. A partial evaluation can turn out be of no more use than no evaluation at all. Remember that most meaningful black box metrics at the dialogue level cannot be automated, given the current state of the art.
- 6. If the system to be evaluated is intended for use in a real context, ensure that the test conditions match the end-use conditions as closely as possible.
- 7. Wherever possible, use evaluation metrics which have already been described in the literature.
- 8. Where it is necessary to invent some new metric ensure (i) that it is wellmotivated, (ii) that it is fair (not favouring your system unreasonably), and (iii) that it is fully described whenever it is mentioned in public documents.
- 9. Be very cautious when comparing systems. Valid conclusions may not be drawn when significant differences exist between (i) the application domains, (ii) the test conditions, and (iii) the metrics used.
- 10. Whole systems can only be compared meaningfully with black box metrics, and not glass box metrics.
- 13.4.5 Readings in interactive dialogue system evaluation

There are still relatively few spoken language dialogue systems around, so the literature on their evaluation is fairly thin. The following readings provide a good introduction to the state of the art: Bailleul (1987), Bates et al. (1990), Bates and Ayuso (1991), Boogart et al. (1993), Cookson (1988), Delogu et al. (1993a), Gerbino et al. (1993), Goodine et al. (1992), Hirschman et al. (1990), House et al. (1992), Simpson and Fraser (1993), Smith et al. (1992), Tubach and Doignon (1991), Vergeynst et al. (1993), and Zue et al. (1991).

Bibliographical references

618 Bibliographical references

References

- D. Abercrombie (1967). *Elements of general phonetics*. Edinburgh University Press, Edinburgh.
- A. Aho, B. Kernighan and P. Weinberger (1987). *The AWK programming language.* Addison-Wesley Publishing Company, Reading, Mass., etc.
- W. Ainsworth (1988). Speech recognition by machine. Peter Peregrinus.
- J. Aitchison (1994). Words in the mind. An introduction to the mental lexicon. Blackwell, Oxford.
- M. Aitkin, D. Anderson, B. Francis and J. Hinde (1989). *Statistical modelling in GLIM*. Clarendon Press, Oxford.
- G. Akers and M. Lennig (1985). Intonation in text-to-speech synthesis: Evaluation of algorithms. Journal of the Acoustical Society of America, JASA 77: 2157– 2165.
- A. Akmajian (1984). Linguistics: An introduction to language and communication. The MIT Press, Cambridge, Massachusetts, 2nd edition.
- G. Allen (1988). The PHONASCII system. Journal of the International Phonetic Association 18(1): 9-25.
- J. Allen, M. Hunnicutt and D. Klatt (1987). From text to speech: The MITalk system. Cambridge University Press, Cambridge.
- M. Allerhand (1987). Knowledge-based speech pattern recognition. Kogan Page, London.
- F. Alleva, H. Hon, X. Huang, M. Hwang, R. Rosenfeld and R. Weide (1992). Applying SPHINX-II to the DARPA Wall Street Journal CSR task. In: Speech and Natural Language workshop, pp. 393–398, Harriman, New York.
- F. Alleva, X. Huang and M.-Y. Hwang (1993). An improved search algorithm using incremental knowledge for continuous speech recognition. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume II, pp. 307–311, Minneapolis, MN, April.
- F. Althoff, G. Drexel, H. Lüngen, M. Pampel and C. Schillo (1996). The treatment of compounds in a morphological component for speech recognition. In: D. Gibbon, ed., Natural language processing and speech technology. Results of the 3rd KONVENS Conference, Bielefeld, October 1996, pp. 71-76. Mouton de Gruyter, Berlin, New York.
- T. Andernach, G. Deville and L. Mortier (1993). The design of a real world Wizard of Oz experiment for a speech driven telephone directory information system.
 In: Proceedings of the 3rd European Conference on Speech Communication and Technology, pp. 1165-1168, Berlin, September.
- F. Andry, E. Bilange, F. Charpentier, K. Choukri, M. Ponamali and S. Soudoplatoff (1990). Computerised simulation tools for the design of an oral dialogue system. In: *Proceedings of the ESPRIT Technical Conference*, Brussels, November.
- F. Andry, S. McGlashan, N. Youd, N. Fraser and S. Thornton (1992). Making DATR work for speech: Lexicon compilation in SUNDIAL. Computational Linguistics 18(3): 245-267.
- J. Argente (1991). From speech to speaking styles. In: Proceedings of the ESCA

Workshop 'Phonetics and phonology of speaking styles: Reduction and elaboration in speech communication', volume 1–1, pp. 1–12, Barcelona.

- B. Atal (1976). Automatic recognition of speakers from their voices. *Proceedings* of the IEEE, April, 64(4): 460.
- B. Atal, J. Miller and R. Kent, eds. (1991). Papers in speech communication: Speech processing. Acoustical Society of America.
- V. Aubergé (1992). Developing a structured lexicon for synthesis of prosody. In:
 G. Bailly, C. Benoît and T. Sawallis, eds., *Talking machines: Theories, models and designs*, pp. 307-321. North-Holland, Amsterdam.
- J. Austin (1962). How to do things with words. Oxford University Press, Oxford.
- D. Autesserre, G. Pérennou and M. Rossi (1989). Methodology for the transcription and labeling of a speech corpus. *Journal of the International Phonetic Association* 19(1): 2-15.
- A. Averbuch, L. Bahl and R. Bakis (1987). Experiments with the TANGORA 20000 word speech recognizer. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 701-704.
- A. Averbuch, L. Bahl, R. Bakis, P. Brown, A. Cole, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. De Souza, E. Epstein, D. Fraleigh, F. Jelinek, S. Katz, B. Lewis, R. Mercer, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman and P. Spinelli (1986). An IBM PC-based large-vocabulary isolated-utterance speech recognizer. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 53–56.
- H. Baayen (1991). De CELEX lexicale databank. Forum der Letteren 32(3): 221-231.
- L. Bahl, P. Brown, P. De Souza and R. Mercer (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions* on Acoustics, Speech and Signal Processing, ASSP-37(7) pp. 1001–1008. Also in: A. Waibel, K.-F. Lee, eds. (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 507–514.
- L. Bahl, F. Jelinek and R. Mercer (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, March*, 5: 179–190.
- L. Bahl, F. Jelinek, R. Mercer and A. Nadas (1984). Next word statistical predictor. IBM Tech. Disclosure Bulletin, December, 27(7A): 3941-3942.
- C. Bailleul (1987). Evaluation des performances d'un système de reconnaissance vocale dans des tâches de contrôle airiens. Note Interne, CENA/N87083, 22 June.
- G. Bailly (1994). Rule compilers and text-to-speech systems. Les Cahiers de l'ICP 3: 87-91.
- G. Bailly and C. Benoît, eds. (1992). Talking machines: Theories, models and designs. North-Holland, Elsevier Science Publishers, Amsterdam.
- J. Baker (1975a). The DRAGON system An overview. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-23 pp. 24-29.
- J. Baker (1975b). Stochastic modeling for automatic speech understanding. In: D. Reddy, ed., Speech recognition, pp. 521-541. Academic Press, New York,
N.Y. Also in: A. Waibel, K.-F. Lee, eds. (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 297–307.

- J. Baker (1989). Dragondictate-30k: Natural language speech recognition with 30000 words. In: Proceedings of the European Conference on Speech Technology, volume 2, pp. 161-163.
- J. Baker, P. Bamberg, K. Bishop, L. Gillick, V. Helman, Z. Huang, Y. Ito, S. Lowe, B. Peskin, R. Roth and F. Scattone (1992). Large vocabulary recognition of Wall Street Journal sentences at Dragon systems. In: Speech and Natural Language Workshop, pp. 387–392, Harriman, New York, 23–26 February.
- M. Ball (1991). Computer coding of the IPA: Extensions to the IPA. Journal of the International Phonetic Association 21(1): 36-41.
- G. Ballou, ed. (1987). Handbook for sound engineers. W. Sams & Co., Indianapolis, U.S.A.
- S. Barber, R. Carlson, P. Cosi, M. Di Benedetto, B. Granström and K. Vagges (1989). A rule-based Italian text-to-speech system. In: *Proceedings of the Eurospeech '89*, volume 2, pp. 517–520, Paris.
- W. Barry and A. Fourcin (1990). Speaker selection criteria. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Interim Report Year I, Reference SAM-UCL-G002, Document SAM-UCL-001.
- W. Barry and A. Fourcin (1992). Levels of labelling. Computer Speech and Language 6: 1-14.
- W. Barry, M. Grice, V. Hazan and A. Fourcin (1989). Excitation distributions for synthesised speech. In: *Proceedings of the Eurospeech '89*, volume 1, pp. 353-356, Paris.
- B. Bartlett (1987). Choosing the right microphones by understanding design tradeoffs. J. Audio. Eng. Soc. 35.
- M. Bates and D. Ayuso (1991). A proposal for incremental dialogue evaluation. In: Proceedings of the DARPA Workshop on Speech and Natural Language, pp. 319-322, Pacific Grove, CA, February.
- M. Bates, S. Boisen and J. Makhoul (1990). Developing an evaluation methodology for spoken language systems. In: Proceedings of the DARPA Workshop on Speech and Natural Language, pp. 102–108, Hidden Valley, PA, June.
- F. Baum (1900). The Wizard of Oz. Collins, London. Edition of 1974.
- L. Baum (1972). An inequality and associated maximization technique in statistical estimation of a Markov process. *Inequalities* 3(1): 1–8.
- M. Beckman (1986). Stress and non-stress accent. Foris, Dordrecht.
- F. Belina and D. Hogrefe (1988). The CCITT specification and design language SDL. Computer networks and ISDN systems 16: 311-341.
- T. Bell, J. Cleary and I. Witten (1990). *Text compression*. Prentice Hall, Englewood Cliffs, NJ.
- C. Benoît (1989). Intelligibility test for the assessment of French synthesizers using semantically unpredictable sentences. In: Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, 1.7.1-1.7.4.
- C. Benoît (1991). On the assessment of audio-visual speech synthesis. In: Pro-

ceedings of the Workshop on International Cooperation and Standardisations of Speech Databases and Speech I/O Assessment Methods, Chiavari, Italy.

- C. Benoît, T. Lallouache, T. Mohamadi and C. Abry (1992). A set of French visemes for visual speech synthesis. In: G. Bailly and C. Benoît, eds., *Talking* machines: Theories, models, and design, pp. 485–504. North Holland, Elsevier Science Publishers, Amsterdam.
- C. Benoît, A. Van Erp, M. Grice, V. Hazan and U. Jekosch (1989). Multilingual synthesizer assessment using semantically unpredictable sentences. In: *Proceedings of the Eurospeech '89*, volume 2, pp. 633-636, Paris.
- P. Bentler (1985). Theory and implementation of EQS, a structural equations program. BMDP Statistical Software Inc., Los Angeles.
- E. Berendsen, S. Langeweg and H. Van Leeuwen (1986). Computational phonology: Merged not mixed. In: Proceedings of the International Conference on Computational Linguistics '86, pp. 612–614.
- A. Berger, P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, J. Gillett, J. Lafferty, R. Mercer, H. Printz and L. Ures (1994). The Candide system for machine translation. In: *Proceedings of the ARPA Human Language Technology Workshop*, pp. 152–157, Plainsboro, NJ, March.
- D. Berkley and J. Flanagan (1990). Integration of speech recognition, text-tospeech synthesis, and talker verification into a hands free audio/image teleconferencing system (humanet). *ICSLP* 20(1): 861-864.
- F. Bimbot, I. Magrin-Chagnolleau and L. Mathan (1995). Second-order statistical measures for text-independent speaker identification. Speech Communication 17. 1–2.
- F. Bimbot and L. Mathan (1993). Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In: *Proceedings of the Eurospeech*, pp. 169–172.
- E. Black, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, R. Roukos, B. Santorini and T. Strazalkowski (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In: *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 306-311, Pacific Grove, CA, February.
- A. Bladon (1990). Evaluating the prosody of text-to-speech synthesizers. In: Proceedings of the Speech Tech '90, pp. 215-220.
- J. Blauert (1983). Spatial hearing. MIT Press, Cambridge.
- D. Bleiching (1992). Prosodisches Wissen im Lexikon. In: G. Görz, ed., KON-VENS 92, 1. Konferenz "Verarbeitung natürlicher Sprache", Nürnberg, 7.-9. Oktober 1992, pp. 59-68. Springer-Verlag, Berlin.
- D. Bleiching, G. Drexel and D. Gibbon (1996). Ein synkretismusmodell für die deutsche morphologie. In: D. Gibbon, ed., Natural language processing and speech technology. Results of the 3rd KONVENS Conference, Bielefeld, October 1996, pp. 237-248. Mouton de Gruyter, Berlin, New York.
- D. Bleiching and D. Gibbon (1994). Handbuch zur Demonstrator-Wortliste. V1.1. May 1994, Bielefeld University, Bielefeld, Germany.

- G. Bloothooft, V. Hazan, D. Huber and J. Llisterri (1995). European studies in phonetics and speech communication. OTS Publications, Utrecht.
- D. Bobrow and T. Winograd (1977). An overview of KRL, a knowledge representation language. *Cognitive Science* 1: 3-46.
- B. Boguraev, J. Carroll, S. Pulman, G. Russell, G. Ritchie, A. Black, E. Briscoe and C. Grover (1988). The lexical component of a natural language toolkit. In: D. Walker, A. Zampolli and N. Calzolari, eds., Automating the lexicon: Research and practice in a multilingual environment. Cambridge University Press, Cambridge.
- D. Bolinger (1972). Accent is predictable (if you're a mind-reader). Language 48: 633-644.
- R. Bolt (1970). Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. JASA 47(2): 597. Part 2.
- T. Boogaart and K. Silverman (1992). Evaluating the overall comprehensibility of speech synthesizers. In: Proceedings of the 2nd International Conference on Spoken Language Processing, ICSLP, pp. 1207-1210, Banff.
- T. Boogart, P. Van Alphen and J. Doll (1993). Application oriented assessment of dialogue systems. In: Joint ESCA – NATO/RSG10 Tutorial and Research Workshop on Applications of Speech Technology, Lautrach, September.
- L. Boves (1984). The phonetic basis of perceptual ratings of running speech. Foris, Dordrecht.
- R. Brachman and H. Levesque (1985). Readings in knowledge representation. Morgan Kaufmann Publishers, Inc., Los Altos, California.
- L. Breiman, J. Friedman, R. Ohlsen and C. Stone (1984). Classification and regression trees. Wadsworth, Belmont, CA.
- J. Bridle, M. Brown and R. Chamberlain (1982). An algorithm for connected word recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 899-902, Paris, May.
- A. Brietzmann, H. Hein, H. Niemann and P. Regel (1983). The Erlangen system for understanding continuous German speech. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 304–307, Boston.
- G. Bristow (1984). Electronic speech synthesis. Collins, London.
- G. Bristow (1986). Electronic speech recognition. Collins, London.
- D. Brouwer and D. De Haan, eds. (1987). Woman's language, socialization and self-image. Foris Publications, Dordrecht.
- C. Browman (1980). Rules for demisyllable synthesis using Lingua, a language interpreter. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 561-564, Denver.
- P. Brown, V. Della Pietra, P. De Souza and R. Mercer (1992). Class-based n-gram models of natural language. *Computational Linguistics* 18(4): 467-479.
- G. Bruce (1989). Report from the IPA Working Group on suprasegmental categories. Working Papers 35, Lund University, Department of Linguistics, Lund pp. 25-40.
- H. Bunt, R.-J. Beun, F. Dols, J. von der Linden and G. thoe Schwartzenberg

(1985). The TENDUM dialogue system and its theoretical basis. *IPO Annual Progress Report* 19: 105–113.

- M. Burrell (1991). Assessment of the degradations of synthetic speech and time frequency warping over different listening levels. In: *Proceedings of the Institute of Acoustics*, volume 13, Pt. 2.
- G. Button (1990). Going up a blind alley: Conflating conversation analysis and computational modelling. In: P. Luff, G. Gilbert and D. Frohlich, eds., Computers and conversation, pp. 67-90. Academic Press, London.
- L. Cahill (1993). Morphonology in the lexicon. In: Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics, pp. 87-96, Utrecht.
- L. Cahill and R. Evans (1990). An application of DATR: The TIC lexicon. In: R. Evans and G. Gazdar, eds., *The DATR Papers*, pp. 31–39. School of Cognitive and Computing Science, University of Sussex, Brighton, 2nd edition.
- J. Campbell (1995). Testing with the YOHO CD-ROM Voice Verification Corpus. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 341-344.
- N. Carbonell and J. Pierrel (1986). Architecture and knowledge sources in a human computer oral dialog system. In: *Proceedings of the NATO workshop: Structure of multimodal dialogues including voice*, Corsica, France.
- R. Carlson, B. Granström and D. Klatt (1979). Some notes on the perception of temporal patterns in speech. In: *Proceedings of the 9th International Congress* of Phonetics Sciences, volume 2, pp. 260-267, Copenhagen.
- J. Carroll and J. Chang (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of the "eckhard-young" composition. *Psychometrika* 35: 283-319.
- J. Carson-Berndsen (1993). Time map phonology and the projection problem in spoken language recognition. Doctoral dissertation, University of Bielefeld, Bielefeld, Germany.
- M. Cartier, F. Emerald, D. Pascal, P. Combescure and A. Soubigou (1992). Une méthode d'évaluation multicritère de sorties vocales: Application au test de 4 systèmes de synthèse à partir du texte. In: 19èmes Journées d'Étude sur la Parole, Brussels.
- CCITT (1988a). Artificial voices. Blue Book IXth Plenary Assembly V: 87–99. Recommendation P.50.
- CCITT (1988b). Objective measurement of active speech level. Rec. P. 56 Melbourne, CCITT.
- W. Chafe (1992). The importance of corpus linguistics to understanding the nature of language. In: J. Svartvik, ed., *Directions in corpus linguistics: Proceedings of the Nobel Symposium 82, New York*, pp. 79–97, Berlin. Mouton de Gruyter.
- E. Charniak and D. McDermott (1985). Introduction to Artificial Intelligence. Addison-Wesley, Reading, Massachusetts.
- G. Chollet and C. Gagnoulet (1981). On the evaluation of recognizers and databases using a reference system. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Atlanta.

- N. Chomsky (1965). Aspects of the theory of syntax. The MIT Press, Cambridge, MA.
- N. Chomsky and M. Halle (1968). The sound pattern of English. Harper and Row, New York, Evanston, London.
- K. Choukri, G. Chollet and C. Montacié (1988). Test workstation for the evaluation of speech recognition algorithms, applications and databases. In: *Proceed*ings of the 7th FASE Symposium (Speech'88), pp. 145–151, Edinburgh, August 1988.
- K. Church (1987a). Phonological parsing and lexical retrieval. Cognition 25: 53-69.
- K. Church (1987b). *Phonological parsing in speech recognition*. Kluwer Academic Publishers, Boston, Dordrecht, Lancaster.
- J. Coates (1986). Women, men and language: A sociolinguistic account of sex differences in language. Longman, London.
- Cole (1995). The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on Speech and Audio Processing* 3: 1–20.
- P. Combescure (1981). 20 listes de dix phrases phonétiquement équilibrées. Revue d'Acoustique 56: 34–38.
- A. Content, P. Mousty and M. Radeau (1990). Brulex, une base de données lexicales informatise pour le français écrit et parlé. L'Année Psychologique 90: 551–566.
- S. Cookson (1988). Final evaluation of VODIS voice operated database inquiry system. In: Proceedings of Speech-88, 7th FASE Symposium, pp. 1311-1320, Edinburgh, August.
- P. Cosi and M. Omologo (1991). Caratterizzazione statistica della segmentazione manuale del segnale vocale. Associazione Italiana Acustica (AIA) Meeting. Napoli, Italy, 10-12 April. Cited in Barry and Fourcin 1992.
- S. Crowdy (1993). Spoken corpus design and transcription. Longman, Harlow.
- D. Cruse (1986). Lexical semantics. CUP, Cambridge.
- D. Crystal (1980). Introduction to language pathology. Edward Arnold Ltd., London.
- D. Crystal (1985). A dictionary of linguistics and phonetics. Basil Blackwell, Oxford, UK.
- C. Cucchiarini (1993). Phonetic transcription: A methodological and empirical study. Doctoral thesis, University of Nijmegen, Nijmegen.
- N. Dahlbäck and A. Jönsson (1986). A system for studying human-computer dialogues in natural language. Research Report LiTH-IDA-R-86-42, Department of Computer and Information Science, Linköping University, Linköping.
- N. Dahlbäck and A. Jönsson (1989). Empirical studies of discourse representations for natural language interfaces. In: *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 291– 298, Manchester.
- P. Dalsgaard and A. Baekgaard (1994). Spoken language dialogue systems. In: H. Niemann, R. De Mori and G. Hanrieder, eds., *Progress and prospects in speech and language technology*, pp. 178–191. Infix, Sankt Augustin.

- M. Damhuis, T. Boogaart, C. in 't Veld, M. Versteijlen, W. Schelvis, L. Bos and L. Boves (1994). Creation & analysis of the Dutch Polyphone Corpus. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP, pp. 1803–1806, Yokohama.
- D. Davis and C. Davis (1975). Sound system engineering. W. Sams & Co., Indianapolis, U.S.A.
- R. De Mori, M. Gilloux, G. Mercier, M. Simon, C. Tarrides and J. Vaissiere (1984). Integration of acoustic, phonetic, prosodic and lexical knowledge in an expert system for speech understanding. In: *Proceedings of the IEEE International* Conference on Acoustics, Speech and Signal Processing, ICASSP. 42.9.1–42.9.4.
- J. De Pijper (1983). Modelling British English intonation. Foris, Dordrecht.
- S. Della Pietra, V. Della Pietra, J. Gillett, J. Lafferty, H. Printz and L. Ures (1994). Inference and estimation of a long-range trigram model. Second International Colloquium 'Grammatical Inference and Applications', Alicante, Spain, September 1994 pp. 78-92. Springer-Verlag, Berlin.
- C. Delogu, A. Di Carlo, C. Sementino and S. Stecconi (1993a). A methodology for evaluating human-machine spoken language interaction. In: *Proceedings of* the 3rd European Conference on Speech Communication and Technology, pp. 1427-1430, Berlin, September.
- C. Delogu, A. Paoloni, P. Pocci and C. Sementina (1991). Quality evaluations of text-to-speech synthesizers using magnitude estimation, categorical estimation, pair comparison and reaction time methods. In: *Proceedings of the Eurospeech* '91, pp. 353-355, Genova.
- C. Delogu, A. Paoloni, P. Ridolfi and K. Vagges (1993b). Intelligibility of Italian text-to-speech synthesizers over ortophonic and telephonic channel. In: *Proceedings of the Eurospeech '93*, volume 3, pp. 1893–1896, Berlin.
- C. Delogu, A. Paoloni and C. Sementina (1992a). Comprehension of natural and synthetic speech: Preliminary studies. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Final report, Year three, 1.III.91– 28.II.1992. SAM Internal Report II.c.
- C. Delogu, P. Paoloni, P. Pocci and C. Sementina (1992b). A comparison among different methodologies for evaluating the quality of text-to-speech synthesis systems. In: ESPRIT Project 2589 (SAM), ed., *Multilingual speech input/output assessment, methodology and standardisation*. University College London, London. Final report, Year three, 1.III.91-28.II.1992. SAM Internal Report II.d.
- D. Delomier, A. Meunier and M.-A. Morel (1989). Linguistic features of humanmachine oral interaction. In: *Proceedings of the Eurospeech '89*, volume 2, pp. 236-239, Paris.
- A. Dempster, M. Laird and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. Ser. B (methodological) 39: 1-38.
- E. Den Os (1994). Transliteration of the Dutch Speech Styles Corpus. In: Proceedings of the Institute of Phonetic Sciences, volume 18, pp. 87–94, University of Amsterdam.

- A. Derouault and B. Merialdo (1986). Natural language modelling for phonemeto-text transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence, November*, 8: 742–749.
- D. Diaper (1986). Identifying the knowledge requirements of an expert system's natural language processing interface. In: M. Harrison and A. Monk, eds., People and Computers V: Proceedings of the 2nd Conference of the British Computer Society Human-Computer Interaction Specialist Group, Cambridge. Cambridge University Press.
- D. Diaper (1989). The Wizard's apprentice: A program to help analyse natural language dialogues. In: A. Sutcliffe and L. Macaulay, eds., People and Computers: Designing for usability. Proceedings of the 2nd Conference of the British Computer Society Human-Computer Interaction Specialist Group, Cambridge. Cambridge University Press.
- G. Doddington (1985). Speaker recognition Identifying people by their voices. Proceedings of the IEEE, November, 73(11): 1651.
- J.-M. Dolmazon, J.-C. Caërou and W. Barry (1990). Initial development of SAM standard workstation. SAM-UCL-022, December, Appendix Se.10, University College London, London.
- D. Dougherty (1990). sed & awk. O'Reilly & Associates Inc., Sebastopol, CA.
- G. Dreckschmidt (1987). The linguistic component in the speech understanding system SPICOS. In: H. Tillmann and G. Willée, eds., Analyse und Synthese gesprochener Sprache, Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung, Bonn, pp. 96–101. Olms, Hildesheim.
- R. Drullman and R. Collier (1993). Speech synthesis with accented and unaccented diphones. In: V. Van Heuven and L. Pols, eds., Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation, pp. 147–156. Mouton de Gruyter, Berlin.
- R. Duda and P. Hart (1973). Pattern classification and scene analysis. J. Wiley, New York.
- S. Duncan (1974). On signalling that it's your turn to speak. Journal of Experimental Social Psychology 10: 234-247.
- H. Dybkjaer, N. Bernsen and L. Dybkjaer (1993). Wizard-of-Oz and the tradeoff between naturalness and recognizer constraints. In: Proceedings of the 3rd European Conference on Speech Communication and Technology, pp. 947–950, Berlin, September.
- J. Eargle (1976). Sound recording. Van Nostrand Reinhold Company, New York, USA.
- J. Edwards and M. Lampert, eds. (1993). Talking data: Transcription and coding in discourse research. Lawrence Erlbaum, Hillsdale.
- B. Efron and R. Tibshirani (1993). An introduction to the bootstrap. Chapman & Hall, New York.
- J. Egan (1948). Articulation testing methods. Laryngoscope 58: 955-991.
- U. Ehrlich (1986). Ein Lexikon für das natürlich-sprachliche Dialogsystem EVAR. Arbeitsberichte des IMMD, vol. 19, University of Erlangen-Nürnberg, Erlangen, Germany.

- B. Eisen (1993). Reliability of speech segmentation and labelling at different levels of transcription. In: Proceedings of the Third European Conference on Speech Communication and Technology, volume 1, pp. 673-676, 21-23 September 1993, Berlin, Germany.
- L. Erman (1977). A functional description of the HEARSAY-II speech understanding system. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Hartford.
- L. Erman and F. Hayes-Roth (1981). The HEARSAY-II speech understanding system: Integrating knowledge to resolve uncertainty. In: B. Webber and N. Nilsson, eds., *Readings in Artificial Intelligence*, pp. 349–389. Tioga, Palo Alto, CA.
- L. Erman and V. Lesser (1980). The HEARSAY-II speech understanding system: A tutorial. In: W. Lea, ed., *Trends in speech recognition*, pp. 361–381. Prentice Hall, Englewood Cliffs, NJ. Also in: A. Waibel and K.-F. Lee, eds. (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 235–245.
- J. Esling (1988). 7.1 Computer coding of IPA symbols and 7.3 detailed phonetic representation of computer data bases. *Journal of the International Phonetic Association* 18(2): 99–106.
- J. Esling (1990). Computer coding of the IPA: Supplementary report. Journal of the International Phonetic Association 20(1): 22-26.
- J. Esling and H. Gaylord (1993). Computer codes for phonetic symbols. Journal of the International Phonetic Association 23(2): 83-97.
- U. Essen and V. Steinbiss (1992). Cooccurrence smoothing for stochastic language modelling. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume I, pp. 161–164, San Francisco, CA, March.
- R. Evans and G. Gazdar (1989). The DATR papers. Research Report: May 1989, School of Cognitive and Computing Science, University of Sussex, School of Cognitive and Computing Science, University of Sussex, Brighton.
- R. Evans and G. Gazdar (1990). The DATR papers. Research Report: February 1990, School of Cognitive and Computing Science, University of Sussex, School of Cognitive and Computing Science, University of Sussex, Brighton.
- A. Federico (1989). Comparison between automatic methods and human listeners in speaker recognition tasks. In: *Proceedings of the Eurospeech*, pp. 279–282.
- K. Fellbaum, H. Klaus and J. Sotscheck (1994). Hörversuche zur Beurteilung der Sprachqualität von Sprachsynthesesystemen für die deutsche Sprache. In: Fortschritte der Akustik, Plenarvorträge und Fachbeiträge der 20. Deutschen Jahrestagung für Akustik, pp. 117–122, Dresden, DPG GmbH.
- G. Ferguson (1976). Statistical analysis in psychology and education. McGraw-Hill, Tokyo.
- I. Ferrané, M. De Calmès, D. Cotto, J.-M. Pécatte and G. Pérennou (1992). Statistiques lexicales sur le corpus de textes utilisés dans le projet BREF: Questions de couverture lexicale. In: *Proceedings Communication Homme-Machine*, *Séminaire LEXIQUE*, pp. 217-226, 21-22 January 1992, IRIT-UPS, Toulouse.

- C. Fillmore (1968). The case for case. In: E. Bach and R. Harms, eds., Universals in linguistic theory, pp. 1–88. Holt, Rinehart and Winston, New York.
- L. Fissore, E. Giachin, P. Laface and P. Massafra (1993). Using grammars in forward and backward search. In: *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1525–1528, Berlin, September.
- J. Flanagan, D. Berkley, G. Elko and M. Sondhi (1991). Autodirective microphone systems. Acoustica 73: 58–71.
- A. Fourcin (1993). The SAM project. Ellis Horwood, Chichester.
- A. Fourcin, G. Harland, W. Barry and V. Hazan, eds. (1989). Speech input and output assessment. Multilingual methods and standards. Ellis Horwood Ltd., Chichester.
- N. Fraser (1991). Corpus-based evaluation of the SUNDIAL system. In: J. Neal and S. Walter, eds., *Proceedings of the Natural Language Processing Systems Evaluation Workshop*, Rome. Rome Laboratory. Technical Report RL-TR-91-362.
- N. Fraser and G. Gilbert (1991a). Effects of system voice quality on user utterances in speech dialogue systems. In: Proceedings of the Second European Conference on Speech Communication and Technology, pp. 57-60, Genova, September.
- N. Fraser and G. Gilbert (1991b). Simulating speech systems. Computer Speech and Language 5: 81-99.
- N. Fraser, N. Gilbert and C. McDermid (1992). The value of simulation data. In: Proceedings of the Workshop on Empirical Models and Methodology for Natural Language Dialogue Systems, Trento, April.
- J. French (1991). Updated notes for soundprint transcribers + one page sample text from COBUILD corpus. Working paper, NERC-WP4-47, October, J.P. French Associated, York and COBUILD, Birmingham.
- J. French (1992). Transcription proposals: Multi-level system. Working paper, NERC-WP 4-50, October, University of Birmigham, Birmingham.
- K. Fu (1982). Syntactic pattern recognition and applications. Prentice-Hall, Englewood Cliffs, NJ.
- S. Furui (1981). Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing 29(2).
- S. Furui (1994). An overview of speaker verification technology. In: ESCA-ETRW Workshop, pp. 1–10, Martigny.
- M. Generet, H. Ney and F. Wessel (1995). Extensions of absolute discounting for language modelling. In: *Proceedings of the Fourth European Conference on Speech Communication and Technology*, pp. 1245–1248, Madrid, September.
- E. Gerbino, P. Baggia, A. Ciaramella and C. Rullent (1993). Test and evaluation of a spoken dialogue system. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP'93, Minneapolis, April.
- P. Geutner (1995). Using morphology towards better large-vocabulary speech recognition systems. Interactive Systems Laboratories, University of Karlsruhe, Karlsruhe, Germany.
- D. Gibbon (1991). Lexical signs and lexicon structure: Phonology and prosody in the ASL-lexicon. Research Report ASL-MEMO-20-91/UBI, University of

Bielefeld, Bielefeld, Germany.

- D. Gibbon (1992a). ILEX: A linguistic approach to computational lexica. In: U. Klenk, ed., Computatio linguae. Aufsätze zur algorithmischen und quantitativen Analyse der Sprache, pp. 32-51. Franz Steiner Verlag, Stuttgart.
- D. Gibbon (1992b). Language and software, or: Fritzl's quest. In: C. Floyd, H. Züllighoven, R. Budde and R. Keil-Slavik, eds., Software Development and Reality Construction, pp. 376–390. Springer Verlag, Berlin, Heidelberg, New York.
- D. Gibbon (1993). Generalized DATR for flexible lexical access: PROLOG specification. VERBMOBIL Report 2, October 1993, Bielefeld University, Bielefeld, Germany.
- D. Gibbon (1995). The VERBMOBIL lexicon: Bielefeld lexicon database V2.1. VERBMOBIL Technisches Dokument 21, 31 January 1995, Bielefeld University, Bielefeld, Germany.
- D. Gibbon and U. Ehrlich (1995). Spezifikationen für ein VERBMOBIL-Lexikondatenbankkonzept. VERBMOBIL Memo 69, Bielefeld University & Daimler Benz AG, Bielefeld, Ulm.
- H. Gilbert and G. Weismer (1974). The effect of smoking on the speaking fundamental frequency of adult women. Journal of Psycholinguistic Research 3: 225-231.
- H. Gish, M. Kraner, W. Russel and J. Wolf (1986). Methods and experiments for text-independent speaker recognition over the telephone line. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, p. 865. 17.2.1.
- H. Gish and M. Schmidt (1994). Text-independent speaker identification. In: *IEEE Signal Processing*, volume 11, pp. 18–32.
- J. Goldsmith (1990). Autosegmental and metrical phonology. Indiana University Linguistics Club, Bloomington, Indiana.
- M. Goldstein (1995). Classification of methods used for assessment of text-tospeech systems according to the demands placed on the listener. Speech Communication 16: 225-244.
- M. Goldstein, B. Lindström and O. Till (1992). Assessing global performance of speech synthesizers: Context effects when assessing naturalness of Swedish sentence-pairs generated by 4 systems using 3 different assessment procedures (free number magnitude estimation, 5- and 11-point category scales). In: ES-PRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. SAM Internal Report II.a, Final report, Year three: 1.III.91-28.II.1992.
- M. Goldstein and O. Till (1992). Assessing segmental intelligibility of two rulebased synthesizers and natural speech using the ESPRIT/SAMVCV test procedures (SOAP v3.0) in Swedish and testing for differences between two correlated proportions. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. SAM Internal Report II.b, Final report, Year three: 1.III.91-28.II.1992.
- Y. Gong (1995). Speech recognition in noisy environments: A survey. Speech

Communication 16: 261–291.

- R. Gonzalez and M. Thomason (1978). Syntactic pattern recognition: An introduction. Addison-Wesley, Reading, MA.
- I. Good (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *December*, 40: 237-264.
- D. Goodine, L. Hirschman, J. Polifroni, S. Seneff and V. Zue (1992). Evaluating interactive spoken language systems. In: *Proceedings of the International Conference on Spoken Language Processing*, *ICSLP'92*, pp. 201–204, Banff, October.
- L. Goorfin (1989). Electronic dictionary pronounces over 83,000 words. Speech Technology 4(4): 49-51.
- A. Gorin, S. Levinson, A. Gertner and E. Goldman (1991). Adaptive acquisition of language. Computer, Speech and Language, April, 5(2): 101-132.
- C. Gray and G. Kopp (1944). Voiceprint identification. Bell Telephone Report, Bell Laboratories.
- D. Green (1986). Control, activation and resource: A framework and a model for the control of speech in bilinguals. *Brain and Language* 27: 210–223.
- S. Greenspan, H. Nusbaum and D. Pisoni (1985). Perception of speech generated by rule: Effects of training and attentional limitations. Research on Speech Perception Progress Report 11, pages 263–287, Indiana University, Indianapolis.
- Y. Grenier (1977). Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique. Ph.D. Thesis.
- H. Grice (1975). Logic and conversation. In: P. Cole and J. Morgan, eds., Syntax and semantics 3: Pragmatics, pp. 41–58. Academic Press, New York.
- M. Grice, K. Vagges and D. Hirst (1991). Assessment of intonation in text-tospeech synthesis systems – A pilot test in English and Italian. In: *Proceedings* of the Eurospeech '91, volume 2, pp. 879–882, Genova.
- M. Grice, K. Vagges and D. Hirst (1992a). Prosodic form tests. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Final report, Year three, 1.III.91-28.II.1992, Stage report So. 5, Part One.
- M. Grice, K. Vagges and D. Hirst (1992b). Prosodic function tests. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Final report, Year three, 1.III.91-28.II.1992, Stage report So. 5, Part Two.
- B. Grosz (1977). The representation and use of focus in dialogue understanding. University of California.
- R. Guindon (1988). A multidisciplinary perspective on dialogue structure in useradvisor dialogues. In: R. Guindon, ed., Cognitive Science and its applications for human-computer interaction, pp. 163-200.
- R. Guindon, K. Shuldberg and J. Connor (1987). Grammatical and ungrammatical structures in user-advisor dialogues: Evidence for sufficiency of restricted languages in natural language interfaces to advisory systems. In: *Proceedings* of the 25th Annual Meeting of the Association for Computational Linguistics, pp. 41-44, Stanford.

- R. Guindon, P. Sladky, H. Brunner and J. Connor (1986). The structure of useradviser dialogues: Is there method in their madness? In: Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, pp. 224–230.
- M. Guyomard and J. Siroux (1986a). PALABRE Phase 1 experimental protocol. Technical Report. CNET/TSS/RCP WP4 task 3, April.
- M. Guyomard and J. Siroux (1986b). PALABRE Phase 2 experimental protocol. Technical Report. CNET/TSS/RCP WP4 task 3, May.
- M. Guyomard and J. Siroux (1987). Experimentation in the specification of an oral dialogue. In: H. Niemann, M. Lang and G. Sagerer, eds., Recent Advances in Speech Understanding and Dialog Systems. NATO ASI Series. Series F: Computer and Systems Sciences, Vol. 46, pp. 497-501. Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo.
- M. Guyomard and J. Siroux (1988). Constitution incrementale d'un corpus de dialogues oraux cooperatifs. *Journal Acoustique* 1.
- R. Haeb-Umbach and H. Ney (1994). Improvements in time-synchronous beam search for 10000-word continuous speech recognition. *IEEE Transactions on Speech and Audio Processing, April,* 2: 353-356.
- J. Hansen, C. Pelaez, L. Solana and P. Vossen (1992). Performance assessment and evaluation: Specification document. SUNSTAR Report II.4.
- A. Hauptmann and A. Rudnicky (1988). Talking to computers: An empirical investigation. International Journal of Man-Machine Studies 28: 583-604.
- W. Hayes (1963). Statistics. Holt, Rinehart and Winston, Inc., New York.
- V. Hazan and M. Grice (1989). The assessment of synthetic speech intelligibility using semantically unpredictable sentences. In: *Proceedings of the ESCA Work*shop on Speech Input/Output Assessment and Speech Databases, pp. 1.6.1–1.6.4.
- V. Hazan and B. Shi (1993). Individual variability in the perception of synthetic speech. In: *Proceedings of the Eurospeech '93*, volume 3, pp. 1849–1852, Berlin.
- J. Heemskerk and V. Van Heuven (1993). MORPA, a morpheme lexicon based morphological parser. In: V. Van Heuven and L. Pols, eds., Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation, pp. 67-85. Mouton de Gruyter, Berlin.
- H. Helfrich (1979). Age markers in speech. In: K. Scherer and H. Giles, eds., Social markers in speech, pp. 63-107. Cambridge University Press, Cambridge.
- S. Hertz, J. Kadin and K. Karplus (1985). The DELTA rule development system for speech synthesis from text. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 1589–1601.
- W. Hess (1983). Pitch determination of speech signals. Springer-Verlag, Heidelberg, F.R.G.
- W. Hess, K. Kohler and H. Tillmann (1995). The PhonDat/Verbmobil Speech Corpus. In: *Proceedings of the Eurospeech 95*, Madrid.
- G. Heyer, K. Waldhur and H. Khatchadourian (1991). Motivation, goals and milestones of ESPRIT II MULTILEX. In: *Génie Linguistique 91*, volume 1, Versailles, France, 16–17 January.
- J. Hieronymus, H. Alexander, C. Bennett, I. Cohen, D. Davies, J. Dalby, J. Laver, W. Barry, A. Fourcin and J. Wells (1990). Proposed speech segmentation cri-

teria for the SCRIBE project. SCRIBE Project Report.

- L. Hirschman, D. Dahl, D. McKay, L. Norton and M. Linebarger (1990). Beyond class A: A proposal for automatic evaluation of discourse. In: *Proceedings of the DARPA Workshop on Speech and Natural Language*, pp. 109–112, Hidden Valley, PA, June.
- E. Hjelmquist, B. Jansson and G. Torell (1987). Psychological aspects on blind people's reading of radio-distributed daily newspapers. In: B. Knave and P. Widebäck, eds., Work with display units 86, pp. 187-201. North-Holland, Elsevier Science Publishers, Amsterdam.
- C. Hockett (1958). A course in modern linguistics. Macmillan, New York.
- H. Höge, E. Marschall, O. Schmidbauer and R. Sommer (1985). Worthypothesengenerierung im Projekt SPICOS. In: H. Niemann, ed., Mustererkennung 85, 7. DAGM-Symposium Erlangen, Informatik-Fachberichte, vol. 107, pp. 175–179. Springer-Verlag, Berlin.
- J. Holmes (1988). Speech synthesis and recognition. Van Nostrand Reinhold (UK) Co. Ltd., Wokingham.
- M. Homayounpour, J. Goldman, G. Chollet and J. Vaissiere (1993). Performance comparison of machine and human speaker verification. In: *Proceedings of the Eurospeech*, p. 2295.
- A. House (1988). The recognition of search by machine A bibliography. Academic Press Ltd., New York, N.Y.
- A. House, C. Williams, M. Hecker and K. Kryter (1965). Articulation testing methods: Consonantal differentiation with a closed response set. *Journal of the Acoustical Society of America*, JASA 37: 158–166.
- J. House, Y. Shitara, M. Grice and P. Howard-Jones (1992). Evaluation of prosody in dialogue synthesis. *Speech, Hearing and Language* 6: 89-108.
- T. Houtgast and J. Verhave (1991). A physical approach to speech quality assessment: Correlation patterns in the speech spectrogram. In: Proceedings of the Eurospeech '91, volume 1, pp. 285–288, Genova.
- T. Houtgast and J. Verhave (1992). An objective approach to speech quality. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Stage report So. 9, Final report, Year three: 1.III.91-28.II.1992.
- P. Howard-Jones (1992a). SOAP, Speech Output Assessment Package. Version 4.0, ESPRIT SAM–UCL–042.
- P. Howard-Jones (1992b). Specification of listener dimensions. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Stage report So. 8, Part One, Final report, Year three: 1.III.91-28.II.1992.
- P. Howell (1990). Clear speech and turn-taking cues in telephone dialogue. Report to BT, University College London, London.
- A. Hunt (1991). New commercial applications of telephone-network-based speech recognition and speaker verification. *Proceedings of the Eurospeech* 15(2): 431.
- M. Hunt (1990). Figures of merit for assessing connected-word recognizers. Speech Communication 9: 329-336.

IPDS (1995). CD-ROM#2: The Kiel corpus of spontaneous speech. vol. 1, kiel.

- IPDS (1996). CD-ROM#3: The Kiel corpus of spontaneous speech. vol. 2, kiel.
- ITU-T (1993). Draft recommendation P.8S Subjective performance assessment of the quality of speech voice output devices. Study group 12 – contribution 6, ITU-T.
- R. Jakobson, G. Fant and M. Halle (1951). *Preliminaries to speech analysis*. The MIT Press, Cambridge.
- W. Jassem and P. Lobacz (1989). IPA phonemic transcription using an IBM PC and compatibles. Journal of the International Phonetic Association 19(1): 16-23.
- U. Jekosch (1992). The Cluster-Identification Test. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Internal report II.e, Final report, Year three: 1.III.91-28.II.1992.
- U. Jekosch and L. Pols (1994). A feature-profile for application-specific speech synthesis assessment and devaluation. In: *Proceedings of the 3rd International Conference on Spoken Language Processing, ICSLP*, Yokohama.
- F. Jelinek (1985). A real-time, isolated-word, speech recognition system for dictation transcription. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 858-861.
- F. Jelinek (1991). Self-organized language modeling for speech recognition. In: A. Waibel and K.-F. Lee, eds., *Readings in speech recognition*, pp. 450–506. Morgan Kaufmann Publishers, San Mateo, CA.
- F. Jelinek, J. Lafferty and R. Mercer (1992). Basic methods of probabilistic context free grammars. In: P. Laface and R. De Mori, eds., Speech recognition and understanding, pp. 347-360. Springer, Berlin.
- F. Jelinek and R. Mercer (1980). Interpolated estimation of Markov source parameters from sparse data. In: E. Gelsema and L. Kanal, eds., *Pattern recognition* in practice, pp. 381–397. North-Holland Publishing Company, Amsterdam.
- F. Jelinek, R. Mercer and S. Roukos (1990). Classifying words for improved statistical language models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 621–624, Albuquerque, NM, April.
- F. Jelinek, R. Mercer and S. Roukos (1991a). Principles of lexical language modeling for speech recognition. In: S. Furui and M. Sondhi, eds., Advances in Speech Signal Processing, pp. 651–699. Marcel Dekker, New York.
- F. Jelinek, B. Merialdo, S. Roukos and M. Strauss (1991b). A dynamic language model for speech recognition. In: *Proceedings of the DARPA Workshop 'Speech* and Natural Language Workshop', pp. 293–295, Pacific Grove, CA, February.
- R. Johnston (1993). An on-going series of subjective experiments to assess speech output from text-to-speech systems. Unpublished report to CCITT Study Group, No. 12.
- W. Jongenburger and R. Van Bezooijen (1992). Evaluatie van ELK: Attitudes van de gebruikers, verstaanbaarheid en acceptabiliteit van de spraaksynthese, bruikbaarheid van het zoeksysteem. Stichting Spraaktechnologie, Utrecht.

- A. Jönsson and N. Dalbäck (1988). Talking to your computer is not like talking to your best friend. In: Proceedings of the First Scandinavian Conference on Artificial Intelligence, Tromso, Norway.
- J. Joreskog and D. Sorbom (1984). Lisrel VI. Analysis of linear structural relationships by maximum likelihood, instrument variables, and least squares methods. Scientific software, Mooreville, IN.
- L. Karttunen (1983). KIMMO: A general morphological processor. Texas Linguistic Forum 22: 165–186.
- H. Kasuya, Y. Endo and S. Saliu (1993). Novel acoustic measurements of jitter and shimmer characteristics from pathological voice. In: *Proceedings of the Eurospeech '93*, pp. 1973–1976.
- S. Katz (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing, March*, 35: 400-401.
- J. Kelley (1983a). An empirical methodology for writing user-friendly natural language computer applications. In: *Proceedings of the International Conference* of Computer-Human Interaction, CHI '83.
- J. Kelley (1983b). Natural language and computers: Six steps for writing an easy-to-use computer application. The Johns Hopkins University, Baltimore.
- J. Kelley (1984). An interactive design methodology for user-friendly natural language office information applications. Association for Computing Machinery Transactions on Office Information Systems 2: 26-41.
- J. Kerkhoff, J. Wester and L. Boves (1984). A compiler for implementing the linguistic phase of a text-to-speech conversion system. In: H. Bennis and W. Van Lessen-Kloecke, eds., *Linguistics in The Netherlands 1984*, pp. 111–119. Foris, Dordrecht.
- L. Kersta (1962). Voiceprint infallibility. In: Meeting of Acoust. Soc. Am., Seattle.
- G. Kinsey (1994). Using voice recognition with IVR systems. In: AVIOS conference proceedings, pp. 49–56, San Jose.
- K. Kirchhoff (1996). Phonologisch strukturierte hmms zur automatischen spracherkennung. In: D. Gibbon, ed., Natural language processing and speech technology. Results of the 3rd KONVENS Conference, Bielefeld, October 1996, pp. 55-63. Mouton de Gruyter, Berlin, New York.
- D. Klatt (1976). The linguistics uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America, JASA 59: 1208-1221.
- D. Klatt (1977). Review of the ARPA speech understanding project. Journal of the Acoustical Society of America, JASA 62(6): 1345–1366. Also in: A. Waibel, K.-F. Lee, eds., (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 554–575.
- D. Klatt (1982). The KLATTalk text-to-speech conversion system. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 1589–1592.
- D. Klatt (1987). Review of text-to-speech conversion in English. Journal of the Acoustical Society of America 82: 737-793.

- R. Kneser and H. Ney (1995). Improved backing-off for m-gram language modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, volume I, pp. 49-52, Detroit, MI, May.
- G. Knowles and P. Alderson (1995). Working with speech: The computational analysis of formal British English speech. Longmans, London.
- G. Knowles, L. Taylor and B. Williams (1995). A corpus of formal British English speech. Longmans, London.
- D. Knuth (1973). The art of computer programming 3: Sorting and searching. Addison-Wesley, Reading, Massachusetts.
- K. Kohler, M. Pätzold and A. Simpson (1995). From scenario to segment: The controlled elicitation, transcription, segmentation and labelling of spontaneous speech. Arbeitsberichte (AIPUK) 29, Institut für Phonetik und Digitale Sprachverarbeitung, IPDS, Universität Kiel, Kiel/Germany.
- A. Kornai (1991). Formal phonology. Doctoral dissertation, Stanford University, Stanford.
- K. Koskenniemi (1983). Two-level morphology: A general computational model for word-form recognition and production. University of Helsinki, Department of General Linguistics, Helsinki, Finland.
- V. Kraft and T. Portele (1995). Quality evaluation of five German speech synthesis systems. Acta Acustica 3: 351–365.
- K. Kryter (1962a). Methods for the calculation and use of the Articulation Index. Journal of the Acoustical Society of America, JASA 34: 1689-1697.
- K. Kryter (1962b). Validation of the Articulation Index. Journal of the Acoustical Society of America, JASA 34: 1698-1702.
- R. Kuhn and R. De Mori (1990). A Cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, June, 12: 570-583.
- W. Labov (1972). Sociolinguistic patterns. University of Pennsylvania Press, Pennsylvania.
- W. Labov (1994). Principles of linguistic change. Volume 1: Internal factors. Blackwell, Oxford.
- C. Labrador and P. Dinesh (1984). Experiments in speech interaction with conventional data services. *Interact* '84, pp. 104–108.
- R. Lacouture and Y. Normandin (1993). Efficient lexical access strategies. In: Proceedings of the European Conference on Speech Technology.
- P. Ladefoged (1975). A course in phonetics. Harcourt, Brace, Jovanovich, New York.
- J. Lafferty, D. Sleator and D. Temperley (1992). Grammatical trigrams: A probabilistic model of link grammars. In: *Proceedings of the AAAI Fall Symposium* on *Probabilistic Approaches to Natural Language*, Cambridge, MA.
- H. Langer and D. Gibbon (1992). DATR as a graph representation language for ILEX speech oriented lexica. Research Report, March 1992, ASL-TR-43-92/UBI, University of Bielefeld, Bielefeld, Germany.
- S. Langeweg (1988). The stress system of Dutch. Doctoral dissertation, Leiden University, Leiden.

- 636 Bibliographical references
- D. Larmouth (1986). The legal and ethical status of surreptitious recording in dialect research: Do human subjects guidelines apply? In: D. Larmouth, T. Murray and C. Murray, eds., Legal and ethical issues in surreptitious recording, Publication of the American Dialect Society, number 76. University of Alabama Press, Tuscaloosa and London.
- R. Lau, R. Rosenfeld and S. Roukos (1993). Trigger-based language models: A maximum entropy approach. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, volume II, pp. 45-48, Minneapolis, MN, April.
- J. Laver (1991). The gift of speech. Papers in the analysis of speech and voice, Edinburgh University Press, Edinburgh.
- J. Laver (1994). Principles of phonetics. Cambridge University Press, Cambridge.
- J. Laver, J. McAllister, M. McAllister and M. Jack (1988). A Prolog-based automatic text-to-phoneme conversion system for British English. In: *Proceedings* of the Second Symposium on Advanced Man-Machine Interface through Spoken Language, November 19-22, Hawaii.
- J. Laver, M. McAllister and J. McAllister (1989). Pre-processing of anomalous text-strings in an automatic text-to-speech system. In: S. Ramsaran, ed., *Studies in the pronunciation of English: A commemorative volume in memory of A.C. Gimson.* Croon Helm, London.
- W. Lea, ed. (1980). Trends in speech recognition. Prentice-Hall, Englewood Cliffs, NJ.
- K.-F. Lee, H.-W. Hon and R. Reddy (1990). An overview of the SHPINX speech recognition system. In: A. Waibel and K.-F. Lee, eds., *Readings in speech* recognition, pp. 600-610. Morgan Kaufmann Publishers, San Mateo, California.
- J. Leggett and G. Williams (1984). An empirical investigation of voice as an input modality for computer programming. *International Journal of Man-Machine* Studies 21: 493–520.
- I. Lehiste (1970). Suprasegmentals. MIT Press, Cambridge, Mass.
- I. Lehiste, J. Olive and L. Streeter (1976). Role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America, JASA 60: 1199-1202.
- E. Lehmann (1983). Theory of point estimation. J. Wiley, New York.
- H. Lehnert and F. Giron (1995). Vocal communication in virtual environments.
 In: Conference documentation of Virtual Reality World '95, pp. 279-293, Stuttgart/Germany.
- V. Lesser, R. Fennell, L. Erman and D. Reddy (1975). Organization of the HEARSAY-II speech understanding system. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-23, pp. 11-23.
- J. Levelt (1989). Speaking: From intonation to articulation. ACL-MIT Press Series in Natural Language Processing. Bradford Book – The MIT-Press, Cambridge Massachusetts, London, England.
- S. Levinson, L. Rabiner and M. Sondhi (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech

recognition. The Bell System Technical Journal, April, 62(4): 1035–1074.

- M. Life, M. Lee and J. Long (1988). Assessing the usability of future speech technology: Towards a method. In: Speech '88: 7th FASE Symposium, Edinburgh.
- R. Likert (1932). A technique for the measurement of attitudes. Archives of Psychology 140.
- R. Linggard (1985). *Electronic synthesis of speech*. Cambridge University Press, Cambridge.
- J. Llisterri (1994). Prosody Encoding Survey, Multext LRE Project 62-050.
- J. Llisterri and J. Mariño (1993). Spanish adaptation of SAMPA and automatic phonetic transcription. In: ESPRIT Project 6819 (SAM-A), ed., Speech technology assessment in multilingual applications, Year 1, 1 April 1993-30 September 1993, pp. 1-9. London. SAM-A periodic progress report, Document No: SAM-A/UPC/001/V1.
- J. Logan, B. Greene and D. Pisoni (1989). Measuring the segmental intelligibility of synthetic speech produced by ten text-to-speech systems. *Journal of the Acoustical Society of America*, JASA 86: 566-581.
- J. Logan, D. Pisoni and B. Greene (1985). Measuring the segmental intelligibility of synthetic speech: Results from eight text-to-speech systems. Research on speech perception Progress Report 11, 3–31, Indiana University, Indianapolis.
- H. Loman and L. Boves (1993). Development of rule based synthesis for text-tospeech. In: V. Van Heuven and L. Pols, eds., Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation, pp. 157–168. Mouton de Gruyter, Berlin.
- B. Lowerre and R. Reddy (1980). The HARPY speech understanding system. In: W. Lea, ed., *Trends in speech recognition*, pp. 340–360. Prentice Hall, Englewood Cliffs, NJ. Also in: A. Waibel, K.-F. Lee, eds., (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 576–586.
- P. Luce, T. Feustel and D. Pisoni (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors* 25: 17-32.
- D. Luzzati and F. Neel (1989). Dialogue behaviour induced by machine. In: Proceedings of the Eurospeech '89, volume 2, pp. 601-604, Paris.
- J. Lyons (1977). Semantics. Volumes I and II. Cambridge University Press, Cambridge.
- B. Maassen and D.-J. Povel (1985). The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. *Journal of the Acoustical Society of America, JASA* 78: 877–886.
- C. MacDermid (1993). Features of naive callers' dialogues with a simulated speech understanding and dialogue system. In: *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pp. 955–958, Berlin, September.
- B. MacWhinney (1995). The CHILDES Project: Tools for analyzing talk. Lawrence Erlbaum, Hillsdale, NJ.
- L. Manous, M. Dedina, H. Nusbaum and D. Pisoni (1985). Speeded sentence verification of natural and synthetic speech. Research on Speech Perception Progress Report 11, Indiana University, Indianapolis.

- L. Marascuilo and R. Serlin (1988). Statistical methods for the social, and behavioral sciences. Freeman and company, New York.
- J. Mariani (1989). Recent advances in speech processing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 429-440.
- W. Marslen-Wilson, ed. (1989). Lexical representation and process. The MIT Press, Cambridge, Massachusetts and London, England.
- B. Mérialdo (1988). Multi-level decoding for very-large-size-dictionary speech recognition. IBM Journal of Research and Development 32(2): 169-301.
- D. Michaelis and H. Strube (1995). Orthogonale akustische Stimmgüteparameter zur Stimmtherapiedokumentation. Fortschritte der Akustik DAGA '95 to be printed.
- A. Monaghan and D. Ladd (1989). Evaluating intonation in the CSTR text-tospeech system. In: Proceedings of the ESCA Workshop on Speech I/O Assessment and speech databases, Noordwijkerhout. 3.6.1-3.6.4.
- A. Monaghan and D. Ladd (1990). Symbolic output as the basis for evaluating intonation in text-to-speech systems. Speech Communication 9: 305-314.
- A. Moody (1991). Speaker verification. Internal Report, January 1991, Ensigma Ltd.
- R. Moore (1977). Evaluating speech recognizers. IEEE Transactions on Acoustics, Speech and Signal Processing 25(2): 178–183.
- R. Moore (1986). The NATO research study group on speech processing: RSG10. In: Proceedings of the Speech Tech'86, pp. 201-203, New York, 28-30 April 1986.
- R. Moore (1988). The technology of speech recognition. In: Proceedings of the CCTA/Blenheim-Online Conference on Knowledge Based Systems in Government, Bristol, 8-10 November 1988.
- R. Moore (1991). International coordination of research standards in speech science and technology. In: Proceedings of the ICSLP-90 Workshop on International Coordination of Spoken Language Database and Assessment Techniques for Speech Input/Output, Kobe, Japan, November 1991.
- R. Moore (1992a). Speech recognition: Available assessment methods and needs for standardisation. In: Proceedings of the Workshop on International Cooperation and Standardisation of Spoken Language Databases and Speech I/O Assessment Techniques, Chiavari, Italy, 26-28 September 1992.
- R. Moore (1992b). User needs in speech research. In: Proceedings of the Workshop on European Textual Corpora, Pisa, Italy, 23-26 January 1992.
- R. Moore (1994a). The "Capability Profile". DRA-CSE Research Note DRA CIS CSE1 RN94/08, August 1994, DRA Speech Research Unit, Malvern, Worcs., UK.
- R. Moore (1994b). The EAGLES working group on spoken language, Advanced Speech Applications. European research on speech technology. In: K. Varghese, S. Pfleger and J. Lefevre, eds., *Research Reports ESPRIT Volume 1.* Springer-Verlag, Berlin.
- S. Mori, C. Suen and K. Yamamoto (1992). Historical review of OCR research

and development. Proceedings of the IEEE, July, 80(7): 1029-1058.

- T. Morimoto, K. Shikano, H. Iida and A. Kurematsu (1990). Integration of speech recognition and language processing in the spoken language translation system SL-TRANS. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP, pp. 921–928, Kyoto.
- E. Moulines and F. Charpentier (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9: 453-467.
- C. Müller and F. Runge (1993). Dialogue design principles key for usability of voice processing. In: Proceedings of the 3rd European Conference on Speech Communication and Technology, pp. 943–946, Berlin, September.
- I. Murray and J. Arnott (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America, JASA* 93: 1097-1108.
- T. Murray and C. Murray (1986). On the legality and ethics of surreptitious recording. In: D. Larmouth, T. Murray and C. Murray, eds., *Legal and ethical issues in surreptitious recording*, Publication of the American Dialect Society, number 76. University of Alabama Press, Tuscaloosa and London.
- H. Murveit, J. Butzberger, V. Digalakis and M. Weintraub (1993). Large vocabulary dictation using SRI's Decipher speech recognition system: Progressive search techniques. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, volume II, pp. 319–322, Minneapolis, MN, April.
- A. Nadas (1984). Estimation of probabilities in the language model of the IBM speech recognition system. IEEE Transactions on Acoustics, Speech and Signal Processing, August, 32: 859-861.
- A. Nadas (1985). On Turing's formula for word probabilities. *IEEE Transactions* on Acoustics, Speech and Signal Processing, December, 33: 1414-1416.
- M. Nespor and I. Vogel (1986). Prosodic phonology. Foris, Dordrecht.
- A. Newell (1978). The palantype transcription unit its history and progress to date. *Hearing*, pp. 99–104. May/June.
- A. Newell (1989). Speech simulation studies performance and dialogue specification. In: J. Peckham, ed., *Recent developments and applications of natural language processing*, pp. 141–157. Kogan Page, London.
- H. Ney (1984). The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing, April,* 32(2): 263-271.
- H. Ney and X. Aubert (1994). A word graph algorithm for large vocabulary, continuous speech recognition. In: *Proceedings of the International Conference on Spoken Language Processing*, *ICSLP*, pp. 1355–1358, Yokohama, Japan, September.
- H. Ney and U. Essen (1993). Estimating small probabilities by leaving-one-out.
 In: Third European Conference on Speech Communication and Technology, pp. 2239-2242, Berlin, September.
- H. Ney, U. Essen and R. Kneser (1994). On structuring probabilistic dependencies

in language modelling. Computer Speech and Language 8: 1-38.

- H. Ney, D. Mergel, A. Noll and A. Paesele (1992). Data driven search organization for continuous speech recognition. *IEEE Transactions on Signal Processing*, *February*, 40(2): 272–281.
- H. Ney, D. Mergel, A. Noll and A. Paeseler (1988). Overview of speech recognition in the SPICOS system. In: H. Niemann, M. Lang and G. Sagerer, eds., *Recent* advances in speech understanding and dialog systems, volume 46 of NATO ASI Series F, pp. 305-310. Springer-Verlag, Berlin.
- H. Niemann, A. Brietzmann, R. Mühlfeld, P. Regel and G. Schukat (1985). The speech understanding and dialog system EVAR. In: R. De Mori and C. Suen, eds., New systems and architectures for automatic speech recognition and synthesis, NATO ASI Series F, vol. 16, pp. 271–302. Springer-Verlag, Berlin.
- H. Niemann, E. Nöth, M. Mast and E. Schukat-Talamazzini (1992). Ein Lexikon für ein natürlich-sprachliches Dialogsystem. In: *Beiträge des ASL-Lexikonworkshops*, pp. 15–18, Wandlitz, 26–27 November. ASL-TR-40– 92/ZSB.
- F. Nolan (1987). The limits of segmental description. In: Proceedings of the Eleventh International Conference of Phonetic Sciences, volume 5, pp. 411– 414, 1–7 August 1987, Tallinn, Estonia.
- S. Nooteboom and J. Kruijt (1987). Accents, focus distribution, and the perceived distribution of given and new information. Journal of the Acoustical Society of America, JASA 82: 1512–1524.
- M. Nossin (1991). Le projet GENELEX: EUREKA pour les dictionnaires génériques. *Génie Linguistique 91*, volume 1. Versailles, France, 16–17 January 1991.
- A. Nunn and V. Van Heuven (1993). MORPHON: Lexicon-based text-to-phoneme conversion and phonological rules. In: V. Van Heuven and L. Pols, eds., Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation, pp. 88-113. Mouton de Gruyter, Berlin.
- H. Nusbaum, S. Greenspan and D. Pisoni (1986). Perceptual attention in monitoring natural and synthetic speech. Research on Speech Perception Progress Report 12, Indiana University, Indianapolis.
- P. Nye and J. Gaitenby (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research, 37/38, pages 169–190.
- P. Nye, F. Ingemann and L. Donald (1975). Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms. Haskins Laboratories Status Report on Speech Research, 41.
- M. Oerder and H. Ney (1993). Word graphs: An efficient interface between continuous speech recognition and language understanding. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, volume II, pp. 119–122, Minneapolis, MN, April.
- J. Oglesby (1994). What's in a number? Moving beyond the equal error rate. To appear in Speech Communication, August 1995. Preliminary version published in Martigny ETRW, pp. 87–90.

- G. Olsen and J. Olsen (1990). User-centered design of Collaborative Technology. Cognitive Science and Machine Intelligence Laboratory. To appear in Organizational Computing 32.
- M. O'Malley and M. Caisse (1987). How to evaluate text-to-speech systems. Speech Technology 3: 66-75.
- J. O'Neill (1975). Measurement of hearing by tests of speech and language. In: S. Singh, ed., Measurement procedures in speech, hearing, and language, pp. 219-252. University Park Press, Baltimore.
- A. Oppenheim (1978). Applications of digital signal processing. Prentice-Hall, Englewood Cliffs, N.J.
- D. O'Shaughnessy (1986). Speaker recognition. IEEE ASSP Magazine, pp. 4-17.
- D. O'Shaughnessy (1987). Speech communiacation human and machine. Addison-Wesley, New York.
- D. Pallet, W. Fisher and J. Garofolo (1990). DARPA ATIS results, June 1990.
 In: Proceedings of the DARPA Workshop on Speech and Natural Language, pp. 114-121, Hidden Valley, PA, June.
- D. Pallett (1985). Performance assessment of automatic speech recognizers. Journal of the National Bureau of Standards 90(5). September-October 1985.
- A. Parducci (1965). Category judgement: A range-frequency model. Psychological Review 72: 407-418.
- C. Pavlovic, M. Rossi and R. Espesser (1990). Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis system. *Journal of the Acoustical Society of America*, JASA 87: 373-381.
- C. Pavlovic, M. Rossi and R. Espesser (1991). Perceived spectral energy distributions for EUROM-0 speech and for some synthetic speech. In: *Proceedings of the 12th International Congress of Phonetic Sciences*, volume 5, pp. 418-421, Aix-en-Provence.
- J. Peckels and M. Rossi (1973). Le test diagnostic par paires minimales. Adaptation au Français du "Diagnostic Rhyme Test" de W.D. Voiers. *Revue* d'Acoustique 27: 245-262.
- J. Peckham (1990). An overview of speaker verification technology and application over the telephone. In: *Proceedings of the Voice System Worldwide*, p. 166.
- J. Peckham (1993). A new generation of spoken dialogue systems: Results and lessons from the SUNDIAL project. In: *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pp. 33-40, Berlin, September.
- J. Peckham and T. Thomas (1990). Recognizer sensitivity analysis: A method for assessing the performance of speech recognizers. *Speech Communication* 9: 317–328.
- G. Pérennou, D. Cotto, M. De Calmès, I. Ferrané, J. Pécatte and J. Tihoni (1991). Composantes phonologique et orthographique de BDLEX. In: *Deuxièmes Journées Nationales du GRECO-PRC Communication Homme-Machine*, pp. 351–362, Toulouse, 29–30 January.
- G. Pérennou, D. Cotto, M. De Calmès, I. Ferrané and J.-M. Pécatte (1992). Le projet BDLEX de base de données lexicales du Français écrit et parlé. In: Proceedings Communication Homme-Machine, Séminaire LEXIQUE, pp. 153-

171, 21-22 January 1992, IRIT-UPS Toulouse.

- G. Pérennou and M. De Calmès (1987). BDLEX lexical data and knowledge base of spoken and written French. In: European Conference on Speech Technology, volume 1, pp. 393-396, Edinburgh.
- G. Pérennou and J. Tihoni (1992). Lexique et phonologie en reconnaissance de la parole. In: Proceedings Communication Homme-Machine, Séminaire LEX-IQUE, pp. 41-57, 21-22 January 1992, IRIT-UPS Toulouse.
- W. Perkins (1977). Speech pathology, an applied behavioral science. The C.V. Mosby Company, Saint Louis.
- S. Philips, S. Stelle and C. Tanz (1987). Language, gender and sex in comparative perspective. Cambridge University Press, Cambridge.
- R. Pieraccini, E. Levin and E. Vidal (1993). Learning how to understand language.
 In: Third European Conference on Speech Communication and Technology, pp. 1407-1412, Berlin, September.
- A. Pierce (1991). Acoustics: An introduction to its physical principles and applications. McGraw Hill, Inc., New York.
- D. Pisoni, B. Greene and H. Nusbaum (1985a). Perception of synthetic speech generated by rule. *Proceedings of the IEEE* 73: 1665–1676.
- D. Pisoni, B. Greene and H. Nusbaum (1985b). Some human factors issues in the perception of synthetic speech. In: *Proceedings Speech Tech* '85, pp. 57-61, New York.
- J. Pitrelli, M. Beckman and J. Hirschberg (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP, 18-22 September 1994, Yokohama, Japan.
- M. Plenat (1991). Vers d'une phonémisation des sigles. In: Deuxièmes journées du GDR-PRC Communication Homme-Machine, EC2 Editeur, pp. 363-371, Toulouse, 29-30 January.
- R. Plomp and A. Mimpen (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 8: 43–52.
- L. Pols (1991). Quality assessment of text-to-speech synthesis-by-rule. In: S. Furui and M. Sondhi, eds., *Advances in speech signal processing*, pp. 387–416. Marcel Dekker Inc., New York.
- L. Pols, J.-P. Lefevre, G. Boxelaar and N. Van Son (1987). Word intelligibility of a rule synthesis system for French. In: *Proceedings of the European Conference* on Speech Technology, volume 1, pp. 179–182, Edinburgh.
- M. Ponamale, E. Bilange, K. Choukri and S. Soudoplatoff (1990). A computeraided approach to the design of an oral dialogue system. In: *Proceedings of Eastern Multiconference*, Nashville.
- T. Portele, B. Heuft, F. Höfer, H. Meyer and W. Hess (1994). A new high quality speech synthesis system for German. In: *Proceedings Yokohama/New Paltz.*
- R. Pratt (1987). Quantifying the performance of text-to-speech synthesizers. Speech Technology, pp. 54-64.
- P. Price (1990). Evaluation of spoken language systems: The ATIS domain. In: Proceedings of the DARPA Workshop on Speech and Natural Language, pp.

91-95, Hidden Valley, PA, June.

- H. Quené (1993). Segment durations and accent as cues to word segmentation in Dutch. Journal of the Acoustical Society of America, JASA 94: 2027-2035.
- L. Rabiner and R. Schafer (1978). Digital processing of speech signals. Prentice-Hall, Englewood Cliffs, N.J.
- A. Radford (1988). Transformational grammar: A first course. CUP, Cambridge.
- J. Ralston, D. Pisoni, S. Lively, B. Greene and J. Mullennix (1991). Comprehension of synthetic speech produced by rule: Word monitoring and sentence-bysentence listening times. *Human Factors* 33: 471–491.
- M. Rayner, H. Alshawi, I. Breton, D. Carter, V. Digalakis, B. Gamback, J. Kaja, J. Karlgren, B. Lyberg, S. Pulman, P. Price and C. Samuelsson (1993). A speech to speech translation system built from standard components. In: *Proceedings* of a Workshop: Human Language Technology, pp. 217–222, Princeton, NJ, 21– 24 March.
- R. Reilly (1987). Ill-formedness and mis-communication in person-machine dialogue. Information and Software Technology 29: 69-74.
- M. Reyelt, M. Grice, R. Benzmüller, J. Mayer and A. Batliner (1996). Prosodische Etikettierung des Deutschen mit ToBI. In: D. Gibbon, ed., Natural language processing and speech technology. Results of the 3rd KONVENS Conference, Bielefeld, October 1996, pp. 144-155. Mouton de Gruyter, Berlin, New York.
- D. Reynolds (1994). Speaker identification and verification using Gaussian mixture speaker models. To appear in Speech Communication, August 1995. Preliminary version published in ETRW Martigny, pp. 27–30.
- M. Richards and K. Underwood (1984a). How should people and computers speak to each other? *Interact* '84, pp. 33–36.
- M. Richards and K. Underwood (1984b). Talking to machines. How are people naturally inclined to speak? In: E. Megaw, ed., *Contemporary Ergonomics*. Taylor and Francis, London.
- G. Ritchie, A. Black, G. Russell and S. Pulman (1992). Computational morphology. The MIT Press, Cambridge, Massachusetts and London.
- P. Roach, G. Knowles, T. Varadi and S. Arnfield (1993). MARSEC: A machinereadable Spoken English corpus. *Journal of the International Phonetic Association* 23(2): 47–53.
- P. Roach, H. Roach, A. Dew and P. Rowlands (1990). Phonetic analysis and the automatic segmentation and labeling of speech sounds. *Journal of the International Phonetic Association* 20(1): 15-21.
- D. Roe and J. Wilpon (1994). Voice communication between humans and machines. National Academy Press, Washington.
- J. Roelofs (1987). Synthetic speech in practice: Acceptance and efficiency. Behaviour and Information Technology 6: 403-410.
- D. Rose (1971). Audiological assessment. Prentice-Hall International, Inc., London.
- A. Rosenberg (1973). Listener performance in speaker verification tasks. IEEE Transactions on Audio Electroacoustic 21: 221-225.
- A. Rosenberg (1976). Automatic speaker verification: A review. Proceedings of

the IEEE, April, 64(4): 475.

- R. Rosenfeld (1994). Adaptive statistical language modeling: A maximum entropy approach. Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. CMU-CS-94-138.
- M. Rossi (1988). Acoustics and electroacoustics. Artech House, Norwood, MA, USA.
- C. Rowden (1992). Speech processing. McGraw-Hill Book Company, London.
- A. Rudnicky, L. Baumeister, K. De Graff and E. Lehmann (1987). The lexical access component of the CMU continuous speech recognition system. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.
- G. Ruske (1985). Demisyllables as processing units for automatic speech recognition and lexical access. In: R. De Mori and C. Suen, eds., New systems and architectures for automatic speech recognition and synthesis, volume 16 of NATO ASI Series F, pp. 593-611. Springer-Verlag, Berlin.
- G. Ruske and T. Schotola (1981). The efficiency of demisyllable segmentation in the recognition of spoken words. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 971–974, Atlanta.
- H. Sacks, E. Schlegloff and G. Jefferson (1974). A simplest systematics for the organization of turn-taking in conversation. *Language* 50: 697–735.
- G. Sagerer (1990). Automatisches Verstehen gesprochener Sprache, volume 74 of Reihe Informatik. Bibliographisches Institut, Mannheim.
- H. Sakoe (1979). Two-level DP matching A dynamic programming-based pattern matching algorithm for connected word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 27: 588-595.
- P. Salza, G. Di Fabbrizio, M. Oreglia, M. Falcone, C. Sementina and C. Delogu (1993). Development of a context dependent methodology for text-to-speech synthesis evaluation in interactive dialogue systems. In: ESPRIT Project 6819 (SAM-A), ed., Speech technology assessment in multilingual applications. London. Report R2, SAM-A Periodic Progress Report. Year 1, 1 April 1993–30 September 1993.
- SAM (1992). Multi-lingual speech input/output assessment, methodology and standardization. ESPRIT project 2589 (SAM), Final report, Year three, 1 III 91-28 II 1992, Ref: SAM-UCL-G004, University College London, London.
- SAM-A (1993). Speech technology assessment in multilingual applications. ES-PRIT Project 6819 (SAM-A), Report No. 2, Year 1, Ref SAM-A/G002.
- P. Scharpff and V. Van Heuven (1988). Effects of pause insertion on the intelligibility of low quality speech. In: Proceedings of the 7th FASE/Speech '88 Symposium, pp. 261-269, Edinburgh.
- K. Scherer and H. Giles, eds. (1979). Social markers in speech. Cambridge University Press, Cambridge.
- M. Schmidt and G. Watson (1991). The evaluation and optimization of automatic speech segmentation. In: Proceedings of the Second European Conference on

Speech Communication and Technology, Eurospeech 91, volume 2, pp. 701-704, 24-26 September 1991, Genova, Italy.

- S. Schröder, G. Sagerer and H. Niemann (1987). Wissensakquisition mit semantischen Netzwerken. In: E. Paulus, ed., Mustererkennung 87, 9. DAGM-Symposium Braunschweig, Informatik-Fachberichte, pp. 305-309. Springer-Verlag, Berlin.
- E. Schukat-Talamazzini (1993). Automatische Spracherkennung. Habilitationsschrift, Erlangen University, Erlangen, Germany.
- E. Schwab, H. Nusbaum and D. Pisoni (1985). Some effects of training on the perception of synthetic speech. *Human Factors* 27(4): 395-408.
- R. Schwartz and S. Austin (1991). A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 701-704, Toronto, May.
- J. Searle (1969). Speech acts: An essay in the philosophy of language. Cambridge University Press, Cambridge.
- J. Searle (1979). *Expression and meaning*. Cambridge University Press, Cambridge.
- P. Sells (1985). Lectures on contemporary syntactic theories: An introduction to Government-Binding theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar. CSLI Center for the Study of Language and Information, Stanford, California.
- S. Siegel (1956). Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York.
- K. Silverman, S. Basson and S. Levas (1990). Evaluating synthesizer performance: Is segmental intelligibility enough? In: Proceedings of the International Conference on Spoken Language Processing, ICSLP, pp. 981-984, Kobe.
- K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg (1992). ToBI: A standard for labeling English prosody. In: *Proceedings of the 1992 International Conference on Spoken Language Processing, ICSLP*, volume 2, pp. 867–870, 12–16 October 1992, Banff, Canada.
- A. Simpson and N. Fraser (1993). Black box and glass box evaluation of the SUNDIAL system. In: Proceedings of the 3rd European Conference on Speech Communication and Technology, pp. 1423-1426, Berlin, September.
- C. Simpson and J. Ruth (1987a). The phonetic discrimination test for speech recognizers: Part I. Speech Technology March/April.
- C. Simpson and J. Ruth (1987b). The phonetic discrimination test for speech recognizers: Part II. Speech Technology October/November.
- T. Skinner, J. Holt and N. Nguyen (1992). Automatic identity confirmation and adaptive solutions. *Speech Technology* pp. 106-111. February 1992.
- P. Smith (1979). Sex markers in speech. In: K. Scherer and H. Giles, eds., Social markers in speech, pp. 109–146. Cambridge University Press, Cambridge.
- R. Smith, D. Hipp and A. Biermann (1992). A dialog control algorithm and its performance. In: Proceedings of the 3rd Conference on Applied Natural

Language Processing, pp. 9-16, Trento, April.

- M. Soclof (1990). A comparison of spontaneous speech and read speech in humanmachine problem solving dialogues. Massachusetts Institute of Technology.
- F. Soong and E.-F. Huang (1991). A Tree-Trellis Fast Search for finding the n-best sentence hypotheses. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 705-708, Toronto, May.
- F. Soong, A. Rosenberg, B. Juang and L. Rabiner (1987). A Vector Quantization approach to speaker recognition. *AT&T Technical Journal* 66. Issue 2.
- C. Sorin (1994). Towards high-quality multilingual text-to-speech. In: *Proceedings of the CRIM/FORWISS workshop*, pp. 53-62, Munich. Also to appear in H. Niemann, ed., Progress and prospects in research and technology, Infix Publishing Company, Sankt Augustin.
- J. Sotscheck (1982). Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgeräte. Der Fernmeldung 36: 1–84.
- C. Sperberg-McQueen and L. Burnard, eds. (1994). Guidelines for electronic text encoding and interchange. TEI P3. Chapter 1 Transcription of Speech. Association for Computational Linguistics, Association for Computers and the Humanities, Association for Literary and Linguistic Computing, Chicago and Oxford.
- M. Spiegel, M. Altom, M. Macchi and K. Wallace (1990). Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. Speech Communication 9: 279–291.
- R. Sproat, J. Hirschberg and D. Yarowsky (1992). A corpus-based synthesizer. In: Proceedings of the 2nd International Conference on Spoken Language Processing, ICSLP, volume 1, pp. 563-566, Banff.
- H. Steeneken (1982). Ontwikkeling en toetsing van een Nederlandstalige Diagnostische Rijmtest voor het testen van spraakcommunicatiekanalen. Rapport IZF 1982–13, IZF, Soesterberg.
- H. Steeneken (1987). Diagnostic information from subjective and objective intelligibility tests. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Dallas.
- H. Steeneken (1989). Objective and diagnostic assessment of (isolated) word recognizers. In: *Proceedings of the European Speech Conference ESCA*, Paris.
- H. Steeneken (1991). RAMOS Recognizer Assessment by means of Manipulation Of Speech applied. In: *Proceedings of the European Speech Conference ESCA*, Genova.
- V. Steinbiss, B.-H. Tran and H. Ney (1994). Improvements in beam search. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP, pp. 2143-2146, Yokohama, Japan, September.
- K. Stevens, C. Williams, J. Carbonell and B. Woods (1968). Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. JASA 44: 1596-1607.
- M. Stubbs (1984). Discourse analysis. The sociolinguistic analysis of natural language. Blackwell, Oxford.

- B. Sundheim (1991). Third message understanding evaluation and conference (MUC-3): Phase 1 status report. In: Proceedings of the DARPA Workshop on Speech and Natural Language, pp. 301-305, Pacific Grove, CA, February.
- A. Syrdal and B. Sciacca (1994). Testing the intelligibility of text-to-speech output with the Diagnostic Pairs Sentence Intelligibility Evaluation. ITD-94-23828A, Technical Memorandum. Submitted to the Journal of the Acoustical Society of America, JASA, AT&T Bell Laboratories.
- J. 't Hart, R. Collier and A. Cohen (1990). A perceptual study of intonation. Cambridge University Press, Cambridge.
- J. Terken (1985). Use and function of accentuation. Some experiments. Doctoral dissertation, Leiden University, Leiden.
- J. Terken (1993). Human and synthetic intonation: A case study. In: V. Van Heuven and L. Pols, eds., Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation, pp. 241–259. Mouton de Gruyter, Berlin.
- J. Terken and R. Collier (1989). Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch. In: Proceedings of the Eurospeech '89, volume 1, pp. 357-359, Paris.
- M. Thielen (1992). Male and female speech. Ph.D. Thesis, University of Amsterdam, Amsterdam.
- N. Thorsen (1980). A study of the perception of sentence intonation Evidence from Danish. Journal of the Acoustical Society of America, JASA 67: 1014– 1030.
- G. Thurmair (1986). Linguistische Analyse im Projekt Spicos. Kleinheubacher Berichte 29.
- M. Tomlinson (1990). Guide to database generation recording protocol. In: ESPRIT Project 2589 (SAM), ed., Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Interim Report Year I, Reference SAM-UCL-G002, Document SAM-RSRE-012.
- O. Tosi, H. Oyer, W. Asbrook, W. Pedrey, C. Nicol and E. Nash (1972). Experiment of voice identification. JASA 51: 2030–2043.
- J. Tubach and P. Doignon (1991). A system for natural spoken language queries: Design, implementation and assessment. In: *Proceedings of the 2nd European Conference on Speech Communication and Technology*, pp. 1473–1476, Genova, September.
- J.-P. Tubach and L.-J. Bok (1985). ZUT Petit dictionnaire français. Institut de Phonitique de Grenoble, avec le concours du CNRS (GRECO Comm. Parlie), Grenoble.
- A. Turing (1950). Computing machinery and intelligence. Mind 59: 433-460.
- V. Valtech, J. Odell, P. Woodland and S. Young (1994). A dynamic network decoder design for large vocabulary speech recognition. In: *Proceedings of the International Conference on Spoken Language Processing*, *ICSLP*, pp. 1351– 1354, Yokohama, Japan, September.
- R. Van Bezooijen (1986). Lay ratings of long-term voice-and-speech characteristics. In: F. Beukema and A. Hulk, eds., *Linguistics in the Netherlands 1986*,

pp. 1–7. Foris, Dordrecht.

- R. Van Bezooijen (1988). Evaluation of two synthesis systems for Dutch Development and applications of intelligibility tests. SPIN-ASSP Report No. 5, Stichting Spraaktechnologie, Utrecht.
- R. Van Bezooijen (1989). Evaluation of the suitability of Dutch text-to-speech conversion for application in a digital daily newspaper. In: Proceedings of the ESCA Workshop Speech I/O Assessment and Speech Databases, pp. 6.3.1-6.3.4, Noordwijkerhout.
- R. Van Bezooijen and W. Jongenburger (1993). Evaluation of an electronic newspaper for the blind in the Netherlands – intelligibility, acceptability, adequacy, and users' attitudes. In: Proceedings of the ESCA Workshop on Speech and Language Technology for Disabled Persons, pp. 195–198, Stockholm.
- R. Van Bezooijen and L. Pols (1987). Evaluation of two synthesis-by-rule systems for Dutch. In: Proceedings of the European Conference on Speech Technology, volume 1, pp. 179-183.
- R. Van Bezooijen and L. Pols (1989). Evaluation of a sentence accentuation algorithm for a Dutch text-to-speech system. In: *Proceedings of the Eurospeech* '89, volume 1, pp. 218-221, Paris.
- R. Van Bezooijen and L. Pols (1990). Evaluating text-to-speech systems: Some methodological aspects. *Speech Communication* 9: 263–270.
- R. Van Bezooijen and L. Pols (1993). Evaluation of text-to-speech conversion for Dutch. In: V. Van Heuven and L. Pols, eds., Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech conversion, pp. 339-360. Mouton de Gruyter, Berlin.
- R. Van Bezooijen and R. Van Hout (1985). Accentedness ratings and phonological variables as measures of variation in pronunciation. *Language and Speech* 28: 129–142.
- B. Van Coile (1989). The DEPES development system for text-to-speech synthesis.
 In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 250-253.
- D. Van Compernolle, J. Smolders, P. Jaspers and T. Hellemans (1991). Speaker clustering for dialectic robustness in speaker independent recognition. In: *Proceedings of Eurospeech '91*, volume 2, pp. 723-726, Genova.
- W. Van Dommelen (1993). Speaker height and weight identification: A reevaluation of some old dates. *Journal of Phonetics* 21: 337–341.
- J. Van Hemert, U. Adriaens-Porzig and L. Adriaens (1987). Speech synthesis in the SPICOS-project. In: H. Tillmann and G. Willée, eds., Analyse und Synthese gesprochener Sprache: Vorträge im Rahmen der Jahrestagung 1987 der Gesellschaft für Linguistische Datenverarbeitung e.V., Bonn, 4-6 March, pp. 34-39. Olms, Hildesheim.
- V. Van Heuven and P. Scharpff (1991). Acceptability of several speech pausing strategies in low quality speech synthesis; interaction with intelligibility. In: *Proceedings of the 12th International Congress of Phonetic Sciences*, pp. 458–461, Aix-en-Provence.
- Y. Van Holsteijn (1993). TextScan: A preprocessing module for automatic text-to-

speech conversion. In: V. Van Heuven and L. Pols, eds., Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation, pp. 27–41. Mouton de Gruyter, Berlin.

- R. Van Hout (1989). De structuur van taalvariatie, een sociolinguistisch onderzoek naar het stadsdialect van nijmegen. Doctoral dissertation, University of Nijmegen, Nijmegen.
- J. Van Santen (1992). Diagnostic perceptual experiments for text-to-speech system evaluation. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP, volume 1, pp. 555-558.
- J. Van Santen (1993). Perceptual experiments for diagnostic testing of text-tospeech systems. *Computer Speech and Language* 7: 49–100.
- J. Van Santen (1994). Using statistics in text-to-speech system construction. In: Proceedings of the ESCA/IEEE Workshop on Speech Synthesis, pp. 240-243, Mohonk NY.
- N. Van Son, L. Pols, S. Sandri and P. Salza (1988). First quality evaluation of a diphone-based speech synthesis system for Italian. In: *Proceedings of the 7th FASE/Speech '88 Symposium*, volume 2, pp. 429-436, Edinburgh.
- N. Vergeynst, K. Edwards, J. Foster and M. Jack (1993). Spoken dialogues for human-computer interaction over the telephone: Complexity measures. In: *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pp. 1415-1418, Berlin, September.
- T. Vintsyuk (1971). Elementwise recognition of continuous speech composed of words from a specified dictionary. *Cybernetics, March-April*, 7: 133-143.
- W. Voiers (1977). Diagnostic evaluation of speech intelligibility. Speech intelligibility and speaker recognition 2: 374–384. Benchmark papers in acoustics, M.E. Hawley (ed.).
- W. Voiers (1983). Evaluating processed speech using the Diagnostic Rhyme Test. Speech Technology 1: 338-352.
- W. Voiers, A. Sharpley and C. Hehmsoth (1975). Research on diagnostic evaluation of speech intelligibility. Research Report AFCRL-72-0694, Air Force Cambridge Research Laboratories, Bedford, Massachusetts.
- J. Vroomen, R. Collier and S. Mozziconacci (1993). Duration and intonation in emotional speech. In: *Proceedings of the Eurospeech '93*, volume 1, pp. 577–580, Berlin.
- W. Wahlster (1993). VERBMOBIL, translation of face-to-face dialogs. In: Proceedings of the Eurospeech '93, opening and plenary sessions, pp. 29-38, Berlin.
- A. Waibel (1988). Prosody and speech recognition. Research notes in artificial intelligence, Pitman Publishing, London.
- A. Waibel, A. Jain, A. McNair, H. Saito, A. Hauptmann and J. Tebelskis (1991). A speech-to-speech translation system using connectionist and symbolic processing strategies. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91, pp. 793-796.
- A. Waibel and K.-F. Lee, eds. (1990). *Readings in speech recognition*. Morgan Kaufmann Publishers, San Mateo, California.
- L. Wall and R. Schwartz (1991). Programming perl. O'Reilly & Associates Inc.,

Sebastopol, CA.

- J. Webers (1985). Tonstudiotechnik. Franzis, Munich, Germany.
- J. Wells (1987). Computer-coded phonetic transcription. Journal of the International Phonetic Association 17(2): 94-114.
- J. Wells (1989). Computer-coded phonemic notation of individual languages of the European Community. Journal of the International Phonetic Association 19(1): 31-54.
- J. Wells (1993a). Applying SAM-PA to Spanish, Portuguese, and Greek: A preliminary discussion document. In: ESPRIT Project 6819 (SAM-A), ed., Speech technology assessment in multilingual applications. London. Document No: SAM-A/D1-Appendix B, SAM-A periodic progress report, Year 1, 1 April 1993-30 September 1993.
- J. Wells (1993b). An update on SAMPA. In: ESPRIT Project 6819 (SAM-A), ed., Speech technology assessment in multilingual applications, pp. 1–6. London. Document No: SAM-A/D1-Appendix A, SAM-A periodic progress report, Year 1, 1 April 1993–30 September 1993.
- S. Whittaker and P. Stenton (1989). User studies and the design of natural language systems. In: Proceedings of the 4th conference of the European Chapter of the Association for Computational Linguistics, pp. 116-123, Manchester.
- N. Willems, R. Collier and J. 't Hart (1988). Synthesis scheme for British English intonation. Journal of the Acoustical Society of America, JASA 84: 1250-1261.
- B. Winer (1971). Statistical principles in experimental design. McGraw-Hill, New York, 2nd edition.
- R. Winski and A. Fourcin (1994). A common European approach to assessment, corpora and standards. In: K. Varghese, S. Pfleger and J. Lefevre, eds., Advanced speech applications. European research on speech technology (Research Reports ESPRIT Volume 1). Springer-Verlag, Berlin.
- R. Winski, R. Moore and D. Gibbon (1995). EAGLES spoken language working group: Overview and results. In: Proceedings of the 4th European Conference on Speech Communication and Technology – Eurospeech'95, pp. 841–844, Madrid, September 1995.
- I. Witten (1982). Principles of computer speech. Academic Press, New York, N.Y.
- P. Woodland, C. Leggetter, J. Odell, V. Valtech and S. Young (1995). The 1994 HTK large vocabulary speech recognition system. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume I, pp. 73-76, Detroit, MI, May.
- W. Woods and V. Zue (1976). Dictionary expansion via phonological rules for a speech understanding system. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 561–564, Philadelphia.
- R. Wooffitt and N. Fraser (1992). We're off to ring the Wizard, the wonderful Wizard of Oz. In: G. Button, ed., *Technology in Working Order: Studies of* work, interaction and technology, pp. 211-230. Routeledge, London.
- M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. Rose, T. Sloboda, M. Tomita, J. Tsutsumi, N. Waibel and W. Ward (1993).

Recent advances in Janus: A speech translation system. In: *Proceedings of a Workshop: Human Language Technology*, pp. 211–216, 21–24 March, Princeton, NJ.

- J. Wright, G. Jones and H. Lloyd-Thomas (1993). A consolidated language model for speech recognition. In: Proceedings of the European Conference on Speech Communication and Technology, pp. 977–980, Berlin, September.
- J. Yamron (1994). A generalization of n-grams. In: *Proceedings of the DARPA* Workshop on Robust Speech Recognition, Rutgers University, Piscataway, NJ, July-August.
- D. Yarrington and R. Foulds (1993). Personalizing synthesized voices. In: Proceedings of the ESCA Workshop on Speech and Language Technologies for Disabled Persons, pp. 169-172, Stockholm.
- S. Young, A. Hauptmann, W. Ward, E. Smith and P. Werner (1989). High level knowledge sources in usable speech recognition systems. *Communications of the ACM* 32(2): 183–194. Also in: A. Waibel and K.-F. Lee, eds., (1990), Readings in speech recognition, Morgan Kaufmann Publishers, San Mateo, California, 538–549.
- V. Zue, J. Glass, D. Goodine, L. Hirschman, H. Leung, M. Phillips, J. Polifroni and S. Seneff (1991). The MIT ATIS system: Preliminary development, spontaneous speech data collection, and performance evaluation. In: *Proceedings of the 2nd European Conference on Speech Communication and Technology*, pp. 537–540, Genova, September.

Part IV

Spoken language reference materials

A Character codes and computer readable alphabets

A.1 Introduction

This appendix discusses the relationship between character sets (or *alphabets*) and their encoding on computers. It is based on the terminology as used in the Unicode standard (Unicode Standard vol. 1.0). Three levels of representation can be discerned:

- character
- glyph
- code

A character is the basic unit of an alphabet. Within the alphabet it has a name, a position, and a content meaning. For example, the character named "a" is the first letter of the standard Latin alphabet. Its content meaning (in the European languages) is loosely related to its pronunciation, i.e. a vocalic sound with the following IPA description: front, low, unrounded. Characters have no visible graphic representation; this representation is produced through *rendering* the character on a suitable medium, e.g. paper, computer screens, etc.

A *glyph* is the "essential shape" of a character; it is the result of the rendering process. Glyphs can be modified through the application of case, font, style, and size operations. For example, the essential shape of the first letter of the Latin alphabet in lower case is the a. In different fonts, this glyph may be modified with a monospaced font to \mathbf{a} , or it may slanted as in a, or boldened as in \mathbf{a} , etc.

A *code* is a mapping of characters to a set of symbols or signs, e.g. numbers, other characters, etc. This mapping is in general an arbitrary one, and it must be known in order to encode a character or decode a code. For example, in 7-bit ASCII, the character "a" is encoded as the 7 bit integer number 97.

A *script* consists of an alphabet and a set of rules that determines the direction of writing (left to right, right to left, up to down, etc.), and the composition of characters (placement of accents, combination of glyphs, etc.).

A.2 ASCII

ASCII codes come in various flavours: the original 7-bit ASCII code, platform dependent variations and extensions such as the Mac ASCII or the country pages of IBM PCs, multinational extensions such as the ISO 8859 family, and application dependent extensions such as ISO 8879 for SGML. 7 bit ASCII

7-bit ASCII (also known as US-ASCII, ANSI X3.4) as defined by the American National Standards Institute is the most widespread code for the computer representation of characters. The 128 numbers of US-ASCII are sufficient for the standard English alphabet, punctuation marks, digits, some mathematical operators, and control codes. However, for many uses, this code system is far too restricted.

The ISO 646 family is a set of standards for 7 bit code tables which differs from US ASCII in language dependent codes, e.g. in the German code table the square brackets and curly braces of the 7-bit ASCII are mapped to German umlauts, in the English code table the # is replaced by a £ symbol, etc.

Platform dependent ASCII

Many hardware vendors, especially in the PC market, implemented proprietary extensions to the 7-bit ASCII standard.

The Macintosh uses an 8-bit ASCII extension which was meant to cover all languages using the Latin alphabet; complex characters could be composed from more than one single character, e.g. by adding an accent or a dieresis. On the IBM PC there exist various 8-bit ASCII extensions for individual languages. This reduces the need for character composition from single characters, but introduces incompatibilities between the different ASCII extensions.

ISO 8859

The International Standards Organization has defined an 8-bit extension to ASCII called the ISO 8859-1 or Latin-1 code. This extension leaves the 7-bit US-ASCII unchanged and adds the most common complex characters from the Latin alphabets. These complex characters include some fractions, special symbols such as the registered Trademark symbol, and accented characters.

The ISO 8859 family has not replaced the platform dependent ASCII codes. However, since it has been officially released as a standard, it serves as a reference code table for the translation of code tables in most forms of electronic communication, e.g. e-mail, news, and others.

ISO 8859 has been extended to non-Latin scripts, e.g. Cyrillic, Arabic, Hebrew, etc. and is renamed according to the languages it is used for, e.g. ISO 8859-5 for Cyrillic, ISO 8859-6 for Arabic, and ISO 8859-8 for Hebrew scripts.

A.2.1 Problems

ASCII and its 8 bit extensions are not sufficient for

- the representation of scripts that contain more than 256 different characters, or
- mixed text documents which use different alphabets in one document.

Ideographic alphabets, e.g. Chinese, Korean, and Japanese, require more than the 256 characters allowed in 8-bit ASCII. Multi-Byte code tables are
needed to adequately represent the character set of these alphabets. Mixed text documents, e.g. regular text with mathematical formulae or phonetic transcriptions, are based on multiple code tables. The document is divided into sections which are encoded with the appropriate code table, and markers assign a code table to each section. In general, switching code tables is achieved through the application of a particular font to a section of the document, e.g. a phonetic font. However, this really is an *abuse* of the font mechanism, because a new glyph is assigned to a particular code, instead of just modifying the original glyph. For example, when the font "Symbol" is applied to a sequence of characters abc on the Macintosh, abc is changed to the characters $\alpha\beta\gamma$ of the Greek alphabet, which is a change of content meaning.

Mixed text documents require that all fonts used in a document be present on the machine the document is processed on. This makes porting to different architectures difficult.

Again, multi-byte character encodings provide sufficient space for all characters needed in a document and thus avoid mixed text documents altogether.

A.3 Multi-Byte encodings

Two multi-byte encodings have been proposed:

- ISO 10646, a four byte (i.e. 2 billion code symbols) code table,
- Unicode standard 1, a two byte (i.e. 65635 code symbols) code table.

The Unicode Standard, Version 1.1, is identical to the two-byte form (UCS-2) of ISO 10646.

A.3.1 ISO 10646

ISO 10646 is a four byte code table that basically assigns sections of the code table to individual alphabets. Characters that occur in more than one alphabet are stored in more than one section – this is especially true for the ideographic alphabets.

To reduce the storage demands for ISO 10646 documents, one- or two-byte subsets of the four-byte code table may be used in a document; control characters are needed to indicate a change of alphabet, e.g. to the four-byte code table.

ISO 10646 has been formally approved in 1992.

A.3.2 The Unicode standard

The Unicode standard is a proposal for a universal two-byte code table for all major written languages. It uses only accepted and official encoding standards for each alphabet to avoid compatibility and acceptance problems. A

¹Unicode is a trademark of Unicode, Inc.

non-profit organization has been founded to promote the Unicode standard:

Unicode, Inc P.O. Box 700519 San Jose, CA 95170-0519 USA

Tel: +1 408 777 5870 Fax: +1 408 777 5082 Email: unicode-inc@unicode.org WWW: http://www.stonehand.com/unicode.html

In the Unicode standard, any glyph is stored only once, and font modifications do not change the essential shape of a glyph. Each glyph has a unique name, number, and content. As a consequence, there no longer exists the notion of a mixed text document.

The code table is divided into sections. The first 256 entries are identical to ISO 8859-1 for compatibility reasons, the other sections contain mathematical symbols, phonetic symbols, non-Latin scripts, vendor specific code tables, and the ideographic alphabets. A rather large section is not standardised, it is reserved for proprietary code tables.

Currently, version 1.1 of the Unicode standard has been published – some 5400 characters from ISO 10646 were added to the code table, and some characters were moved to new locations – and various vendors have announced the support of the Unicode standard (a list of applications that comply with the Unicode standard is available from the Unicode Inc. WWW pages).

A.4 Translation of code tables

The different standard code tables require translator applications that allow the conversion of a document encoded in one code system to a different code system. A translation can either be *total*, i.e. for each character of one alphabet there exists a character in the other alphabet, or *partial*, i.e. there exist characters in one alphabet for which there is no character in the other alphabet. A translation can be *bidirectional* or *unidirectional*. In partial translations, information is lost: either because some characters cannot be represented at all, or because the meaning of a character in one alphabet is different from that in the other alphabet.

A.4.1 Specific vs. general converters

Translator applications are either *specific* or *general* converters. A specific converter translates directly from one code system to another one without

the creation of an intermediate representation. Specific converters are fast, but also restricted in that they can only handle one code table pair -n code tables require n^2 specific converters.

General converters map the source code table to a general representation and then translate from this general representation to the target code table. For a lossless translation, the general representation has to cover all possible source or target code tables; for a lossy translation, partial coverage is sufficient.

General converters require two mapping steps, and the general representation may exceed both the space and time requirements of the source or target code tables.

ISO 8859 and the Unicode standard have shown to be very useful general representations for the translation of character code tables. This is especially true for electronic communication, e.g. mail, news, and the World Wide Web.

A general character code conversion tool is C3, developed at KTH in Sweden. It is in the public domain and can be obtained from

```
http://www.nada.kth.se/i18n/c3/
```

The code tables it can handle include US-ASCII, ISO 8859-X, and Unicode, amongst others.

A.4.2 MIME

MIME stands for "Multipurpose Internet Mail Extensions" and currently is a proposal for a communication encoding standard (RFC 1341). Basically, a MIME header specifies the type of document (text, audio, video, etc.) and the code used to encode the document (ISO 8859, US-ASCII, etc.):

content-type: text/plain; charset=iso-8859-1

Programs with MIME capabilities read the MIME header and either translate the document into a format suitable for the machine it is to be used on, or call external applications to do the translation.

MIME is now supported by most of the news reader, e-mail communication, or World Wide Web viewing applications.

A.5 Computer readable phonetic alphabets

The phonetic alphabet of the IPA is the most commonly used alphabet for the phonetic representation of all spoken languages of the world. It is based on the Latin alphabet, which allowed it to be printed in any better printshop by using the common lead characters.

The code tables for characters in computers have been restricted to 256 entries which could be addressed using one byte, i.e. 8 bits. These code tables are too small to accommodate the full IPA alphabet together with a standard alphabet. Furthermore, IPA characters are often composed of other Latin characters, resulting in a mapping problem: A 1:1 mapping of code table entry to character does not fit into one-byte code tables and requires redundant definitions of font, size, and case operations, whereas composition rules for the combination of base symbols to give an IPA alphabet character are not allowed in many operating systems and thus require proprietary solutions, which make porting a document to another platform almost impossible.

A pragmatic solution to this problem was either to

- restrict the phonetic alphabet to a subset of the IPA alphabet so that it fits into the available space in the code table, or
- create a new nomenclature for the phonetic alphabet which could then be printed like any other text.

The first approach is chosen in SAMPA, where language dependent subsets of the IPA fit into the US-ASCII code table. The second approach is taken in KLATTBET, TIMBET, MRPA, IPA numbers, and others.

With multi-byte alphabets now in the process of standardisation, a direct representation of the IPA alphabet on the level of the operating system has become possible. However, there is still the problem of backwards compatibility, i.e. accessing documents that use a particular one-byte coding system or a proprietary nomenclature.

A.5.1 IPA subsets

The most prominent IPA alphabet is the SAM phonetic alphabet. For each language, a subset of the IPA characters is chosen so that it covers all relevant phonetic categories for that language, and fits into the US-ASCII code table.

SAMPA has been defined for all major European languages. The SAM phonetic alphabets are described in great detail in an appendix of their own (cf. Appendix B).

For American English, the oldest computer readable phonetic alphabet one is the ARPABET. A more detailed one was KLATTBET, used in the MITALK synthesiser (with a space between each phoneme symbol), see Table A.1.

For British English (i.e. RP), the system in use at the CSTR in Edinburgh is the Machine Readable Phonetic Alphabet (MRPA), with a space between each phoneme symbol (see Table A.2).

These coding systems both restrict the alphabet to a subset of the IPA and introduce a new nomenclature for the representation of the phonetic characters.

Symbol	Example	Symbol	Example	Symbol	Example			
PP	pet	TT	ten	KK	core			
BB	bet	DD	debt	GG	gore			
KP	keen	DX	butter	GP	give			
TQ	at Alan	FF	fin	TH	thin			
HH	hat	VV	vat	DH	that			
SS	sat	SH	shin	CH	chin			
ZZ	Z00	ZH	azure	JJ	gin			
MM	met	NN	net	NG	sing			
EN	button	$\mathbf{E}\mathbf{M}$	keep 'em	AXP	(plosive release)			
LL	let	$\mathbf{R}\mathbf{R}$	rent	WW	wet			
YY	yet	\mathbf{EL}	bottle	RX	fire			
ΗX	the hurrah	WH	which	LX	bill			
AA	Bob	AX	about (schwa)	EXR	bear			
IY	beet	UW	boot	AE	bat			
AXR	bar	EY	bait	OW	boat			
UXR	poor	AH	but	AY	bite			
IH	bit	OXR	boar	YU	beauty			
AO	bought	$\mathbf{E}\mathbf{H}$	bet	IX	impunity			
OY	boy	AW	bout	\mathbf{ER}	bird			
IXR	beer	UH	book					
' or 1	primary lexi	cal stress						
" or 2	secondary le	xical stress						
-	syllable boundary							
*	morpheme b	oundary						

Table A.1: The KLATTBET phonetic alphabet

A.5.2 Full IPA alphabet encoding

At the 1989 IPA convention in Kiel a working group was set up to define a coding scheme for the IPA symbols.

The IPA numbers are a radical approach to the problem of representing the IPA character set on computers. While in the other approaches there exists a mnemotechnical relationship between the symbols used in the alphabet and the IPA characters, the IPA numbers are arbitrary and have no obvious relationship with the characters they represent.

According to this scheme, each IPA symbol is uniquely identified through an IPA name and an IPA number in the range of 100 to 999. The range of numbers is divided into classes, e.g. 1nn for consonants, 3nn for vowels, etc. Space is reserved for future extensions and for private use, and even symbols no longer in use are included for backward compatibility (Esling 1990). The IPA encoding has undergone two major revisions, with the current version being that of 1993 (Esling and Gaylord 1993). The IPA code table has become a section of its own in the Unicode and ISO 10646 (UCS-2) standards.

Tables A.5 to A.15 (see below) contain the complete IPA character set

Symbol	Ex.	Symbol	Ex.	Symbol	Ex.	Symbol	Ex.			
р	pet	b	bet	m	met	1	let			
t	tap	d	dock	n	net	r	right			
k	cap	g	gap	ng	sing	w	wing			
h	hang	f	fit	$^{\mathrm{th}}$	thick	s	sat			
$^{\rm sh}$	$_{\rm ship}$	$^{\rm ch}$	chip	v	vote	dh	$_{\mathrm{the}}$			
z	zip	zh	measure	jh	judge	У	young			
i	bit	е	bet	а	bat	0	cot			
uh	cut	u	look	0	around	ii	beet			
aa	cart	00	bought	uu	boot	00	bird			
ai	byte	ei	make	oi	boy	au	cow			
ou	boat	i@	beer	e@	bear	u@	poor			
*	primary lexical stress									
~	secondary lexical stress									
	syllable boundary									

Table A.2: Machine Readable Phonetic Alphabet (MRPA)

ordered by IPA number. They were kindly provided by John Esling (pdb@uvvm.uvic.ca).

The IPA code table is an efficient means of storing phonetic data on computers, while being independent of the character encoding system of any computer platform. Hence, the IPA table, just like the ISO 8859 for ASCII, is often used as a reference standard. Subsets of the IPA, e.g. the national SAM phonetic alphabets, are mapped to IPA symbols, and the translation of one subset to another is possible via the IPA table. IPA numbers have also become a data exchange format for phonetic data since they can be represented in 7-bit ASCII and can thus easily be distributed electronically. The major problem with the IPA code table is that it is not directly accessible by software. In text processing, the font mechanism is often used to substitute Latin glyphs with phonetic glyphs, resulting in the known mixed text document problems. In databases and communication software, IPA numbers are used as an internal representation, but for display and editing purposes this representation has to be mapped to the code tables of the display or the text processing software.

There have been various proposals on using ASCII or ISO 8859 code tables to represent the IPA symbols. These include the systems by

- John Wells of University College London,
- Evan Kirschenbaum of Hewlett-Packard Laboratories,
- David Branner of the University of Washington, and
- James Hieronymus of AT&T Bell Laboratories the Worldbet system.

The system proposed by John Wells (see also Appendix B) is described in a draft report at

http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm

Kirschenbaum's system can be found at

http://alfred1.u.washington.edu:8080/~dillon/ipaascii.html

and a Worldbet description can be retrieved from

ftp://speech.cse.ogi.edu/pub/docs/worldbet.ps.Z

(addresses checked in September 95).

A.6 References

John Esling (1990): "Computer Coding of the IPA: Supplementary Report". Journal of the International Phonetic Association, vol. 20(1).

John Esling and Harry Gaylord (1993): "Computer Codes for Phonetic Symbols". Journal of the International Phonetic Association, vol. 23(2): 83–97.

Unicode Inc. (1991): "The Unicode Standard: Worldwide Character Encoding Version 1.0", Addison Wesley, ISBN 0–201–56788–1

Table A.3: IPA symbol chart

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993, corrected 1996) ONANTS (PULMONIC

COI	SONAN	5 (1 (LIVIC	, (ic)																			
		Bila	abial	Labio	odental	Dent	al	Alve	olar	Post	alveola	r Ret	roflex	Pal	atal	Ve	elar	Uv	ular	Phary	ngeal	G	lottal
Plos	ive	p	b					t	d			t	þ	c	J	k	g	q	G			2	
Nas	al		m		ŋ				n				η		ր		ŋ		Ν				
Tril	I		В						r										R				
Тар	or Flap								ſ				r										
Fric	ative	φ	β	f	v	θ	ð	S	Z	l.	3	ş	Z.	ç	j	Х	γ	χ	R	ħ	ſ	h	ĥ
Late	eral ative		Â					ł	ķ														
App	roximant				υ				ĩ				ł		i		щ						
Late	eral roximant								1				1		λ		L						
	Where sym	ibols a	appear	in pa	irs, the	one to	the	right 1	repre	esents	a voic	ed co	nsonar	nt. Sha	nded a	ireas o	lenote	artic	ulatio	ns judg	ed im	possi	ble.
CO	ISONANT	S (N	ON-P	ULMO	ONIC)								vow	/ELS									
	Cliaba							E						1	Front			Ce	entral			В	ack
$\mathbf{\hat{C}}$	Dilacia		F	oicea	impios	ives	,	E.	jecu	ves			Clos	ie i	•	v –		- i	i • t	ł —		u	• u
ŬŬ	Dental			ואיר היו	labial	aolar	n	, F	sxamı Siləbi	al					/	Ī	Y				υ		
	(Post)alv	eolar	4	P P	ilatal	com	ť	г	Denta	l/alveo	Jar		Clos	e-mid	ı	e.	Ø-		- e -	θ-		γ	•0
· +	Palatoaly	eolar		ſv	alar		k	, ,	/elar	, urreo							Ś			ə			
	Alweeler	lotomi	8	, 1. 1.			, ,		luna	lon fria			Ope	n-mid			8	e cr	<u> </u>	3.0	2	- ^	• •
	Aiveolai	laterai		, .	vuiai		0		iiveo	iai iiic	arre						C	2	0	p			
OTH	ER SYME	BOLS											0	-				æ		р Г	ĺ	a	n
Μ	Voiceless	labial-	velar f	ricative	. (ÇZ	Alve	olo-pa	latal	fricativ	/es		Ope	11		w	here :	c symb	a ● C ols ap	E pear in	pairs,	– u the o	• D ne
W	Voiced lab	bial-ve	lar app	roxima	ıt .	I	Alve	olar la	teral	flap						to	the ri	ght n	eprese	nts a ro	oundec	l vow	/el.
ų	Voiced lat	bial-pa	latal ap	proxim	ant ,	ĥ	Sim	iltaneo	us J	and	Х						SU	PRA	SEGN	AENT.	ALS		
Н	Voiceless	epiglo	tal fric	ative														ī	Prin	narv str	ess		
£	Voiced ep	oiglotta	l fricati	ive		Affricate an be re	es and eprese	l doubl nted b	le arti y two	iculatio 5 symb	ons ols	k	h t	c				i.	Seco	ondary	stress		
	Epiglottal	plosiv	e			joined b	y a ti	e bar it	f nece	essary.		ĸ	P &	9				•	Lon	,fo	ouna et	ˈtɪ]	ən
2140	PITICS	Diac	ritice	may b	e nlace	d abov		vmbo	d wi	th a d	lescend	or o	'n					÷	Half	s -long	e		
	.KIIIC3	Diac	nues d	may 0	e piace	u abov	cas	ymbe h	n wi		escenc	er, e.j	ş. 1 J	+	A			J	Extr	a-short	ĕ		
° ,	/oiceless	I.	i ų t	•	• ^B	reathy v	oicec	<u></u>	2	a o		Dental		<u> </u>	u d	-		Ι	Min	or (foo	t) grou	ıp	
ň	Voiced	<u> </u>	h d	, h	_ C	reaky vo	oiced	<u></u>	?	a 1	u .	Apical		<u><u></u></u>	<u>u</u> 4	_		Ú.	Maj	or (into	natior	ı) gro	oup
	Aspirated	ι 	- u	-	L w	inguolal	bial	<u><u></u></u>	(u ₄w	~	Lamina	d	ľ	<u>a</u>	_		•	Syll	able br	eak .	ii.a	ekt
,	More rounde	a Ç)		" L	abialize	d	- U	" (1" 1i	n	Nasaliz	ed		e 1n	_		$\overline{}$	Linł	cing (al	osence	of a	break)
c	Less rounded	аÇ)		ј _Р	alatalize	:d	U U	, (ր Խ	1	Nasal r	elease		<u>a</u> ¹			TO	VES AT		20.40	CENT	's
÷	Advanced	ų	l		γ v	elarized	l	t	· (18	1	Lateral	release		d ¹	_	"	LEVE	EL Extra	10 110	× C(ONTO	UR
_	Retracted	e			I P	harynge	alized	i t'	0	1']	No aud	ible rele	ease	d	_	é	n -	high		e,₀r	/	Rising
×	Centralized	e			~ v	elarized	l or pl	naryng	ealize	ed 1						_	ē	י ר	Mid		e é	1	High
	Mid-centrali	zed \hat{e}	•		R	aised		ę		(<u>I</u>	= voi	ed alv	eolar fri	icative)			è	י 1	Low		ě	ł	rising Low rising
	Syllabic	ņ	l		r L	owered		ę		_f	= voi	ed bila	ibial app	proxim	ant)		ề	Ŀ	Extra low	ı	è	1	Rising- falling
^	Non-syllabic	ę	2		A	dvanced	l Ton	gue Ro	oot	ę	:						t	Do	ownstep	,	~	Globa	al rise
r	Rhoticity	e	r a	ι	R	etracted	Ton	gue Ro	ot	ę	:						1	U	ostep		7	Globa	al fall

ę

Retracted Tongue Root

≁ Rhoticity

Table A.4: IPA number chart

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993, corrected 1996)

CONSONANT	TS (PULMO	ONIC)								NUMBER	CHART
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	101 102			103 104		105 106	107 108	109 110	111 112		113
Nasal	114	115		116		117	118	119	120		
Trill	121			122					123		
Tap or Flap				124		125					
Fricative	126 127	128 129	130 131	132 133	134 135	136 137	138 139	140 141	142 143	144 145	146 147
Lateral fricative				148 149							
Approximant		150		151		152	153	154			
Lateral approximant				155		156	157	158			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC) VOWELS Clicks Voiced implosives Ejectives Front Centra 176 Bilabial 160 Bilabial 401 Examples: Close 301 309 317 318 - 316 308 401 Examples: 101 Bilabial 103 Dental/alveolar 109 + 401 Velar 132 + 401 Alveolar fricative 177 Dental 162 Dental/alveolar 319 320 321 178 (Post)alveolar 164 Palatal Close-mid 302 310 -- 397 323 - 315 307 179 Palatoalveolar 166 Velar 322 180 Alveolar lateral 168 Uvular Open-mid 303 311 - 326 395 - 314 306 325 OTHER SYMBOLS Open 304 312 -- 305 313 169 Voiceless labial-velar fricative 182 183 Alveolo-palatal fricatives Where symbols appear in pairs, the one to the right represents a rounded vowel. 170 Voiced labial-velar approximant 181 Alveolar lateral flap Simultaneous \int and X171 Voiced labial-palatal approximant 175 SUPRASEGMENTALS 172 Voiceless epiglottal fricative Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. 174 Voiced epiglottal fricative 433 (509) 173 Epiglottal plosive

ę

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. 119 + 402B										
402A	Voiceless	ņ	ģ	405	Breathy voiced	p	a	408	Dental	ţd
403	Voiced	Ş	ţ	406	Creaky voiced	þ	a	409	Apical	ţ₫
404	Aspirated	th	dh	407	Linguolabial	ţ	đ	410	Laminal	ţd
411	More rounded	ş		420	Labialized	tw	dw	424	Nasalized	ẽ
412	Less rounded	ò		421	Palatalized	ť	dj	425	Nasal release	dn
413	Advanced	ų		422	Velarized	t¥	d¥	426	Lateral release	dl
414	Retracted	ē		423	Pharyngealized	t٢	d٢	427	No audible rele	ase d'
415	Centralized	ë		428	Velarized or pha	ryngea	lized 2	09		
416	Mid-centralized	ě		429	Raised	ę	Ĺ	= voic	ed alveolar frica	tive)
431	Syllabic	ņ		430	Lowered	ę	(= voic	ed bilabial appro	ximant)
432	Non-syllabic	ę		417	Advanced Tong	ie Roo	t ę			

419 Rhoticity 327 **A**^{\u03c4} 418 Retracted Tongue Root

501	Primary stress
502	Secondary stress
	,fo∪nə'tı∫ən
503	Long el
504	Half-long e'
505	Extra-short ě
507	Minor (foot) group
508	Major (intonation) group
506	Syllable break Ji.ækt
509	Linking (absence of a break)
TON LEVE	ES AND WORD ACCENTS L CONTOUR

Back

512	519	Extra high	524	529	Rising
513	520	High	525	530	Falling
514	521	Mid	526	531	High rising
515	522	Low	527	532	Low rising
516	523	Extra low	528	533	Rising- falling
517	Down	step	510	Glob	al rise
518	Upste	D	511	Glob	al fall

B SAMPA computer readable phonetic alphabet

B.1 Introduction

SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machinereadable phonetic alphabet originally developed under the ESPRIT project 1541 (SAM) in 1987–89 by an international group of phoneticians and applied in the first instance to Danish, Dutch, English, French, German and Italian (SAM 1988, 1989), later to Norwegian and Swedish (by 1992), and subsequently to Greek, Portuguese, and Spanish (1993). Under the BABEL project it has been extended to Bulgarian, Estonian, Hungarian, Polish, and Romanian.

Section B.2 covers the present status of SAMPA, Section B.3 addresses the individual languages: Bulgarian, Danish, Dutch, English, Estonian, French, German, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Spanish, Swedish, and Greek. Section B.4 discusses additional levels of annotation, and Section B.5 tables the SAMPROSA extension to SAMPA for prosodic annotation.

B.1.1 Notation issues

As with any phonetic notation system, those who developed and applied SAMPA had to make decisions on issues of two types: transcription and coding (see discussion in Wells 1987). The first involves the selection of an appropriate phonetic symbol set; the second involves the allocation of an ASCII number to each symbol that we need, and therefore of a corresponding visual symbol chosen from the character set available on computers.

B.1.2 Transcription

Transcription involves many issues of principle over which phoneticians and linguists have debated for decades. These issues may be new, though, to many engineers and speech technologists. Among such issues are (i) whether the notation should be phonemic, or to some extent allophonic; and if phonemic, how the phoneme set is to be established; (ii) to what extent phonetic symbols should be required to have the same meaning across different languages; and (iii), the relation between the basic, lexical, pronunciation of a word and its actual pronunciation in context.

In principle, SAMPA provides for phonemic notation of languages. For example, the r-sounds of English rip, trip, and drip are all instances of the phoneme /r/, although different articulatory and acoustically (in voicing and in presence/absence of friction). These different allophones are pre-

dictable from the phonetic context: we can unambiguously write them all as /r/. The arguments for preferring phonemic notation to allophonic are (i) it is simpler while still being unambiguous; (ii) correct identification of allophones may be difficult for those without phonetic training; and (iii) too few codes are available in the range 32-127 to provide for all allophones.

In syllable-initial position, English /t/ is alveolar and aspirated; French /t/, dental and unaspirated; Swedish /t/, dental and aspirated. We ignore these comparative differences in our notation, writing all as /t/. SAMPA does not need to adopt distinct symbols to reflect these differences. (However, if and when SAMPA is applied to Hindi, for example, where these differences are phonemic, it would become necessary to notate them explicitly.)

In continuous speech the actual sounds used in pronouncing a word may well differ from the word's citation form (dictionary entry). A phonotypical transcription is one in which citation forms are modified in accordance with known phonetic rules of connected speech. For example, in a phonotypical transcription of English, final linking /r/ would be shown before a following vowel (*better ask*) but not before a consonant (*better go*); the lexical entry would be invariant. In an actual utterance the speaker might or might not conform to phonotypical expectations; an impressionistic transcription reflects a human (or mechanical) auditory or acoustic analysis of what was actually said. In the case at issue, /r/ would be shown if phonetically present in a given instance, not otherwise.

In practice, colleagues working on the various languages to which SAMPA has been applied have chosen to deviate in various respects from these principles. English has plosive /d/ and fricative /ð/ (SAMPA /D/) as distinct phonemes (*den, then*). In Spanish, they are undoubtedly allophones of the same phoneme, and could unambiguously both be written /d/; but for speech technology work our Spanish colleagues prefer to notate them distinctively, as "d" and "D" respectively. The r-sounds in French *rouge, lettre* are different from all the English r-sounds, being respectively a voiced and voiceless uvular fricative. It would seem unambiguous and logical to write them, too, as /r/. But our French colleagues have preferred to use the distinct uvular-r symbol, also provided in SAMPA, namely /R/.

Nevertheless I believe we should as far as possible discourage allophonic and comparative notation. Bulgarian has the simple 6-vowel system, IPA /i e a o u ∂ /. A colleague in Bulgaria has proposed that they be represented in SAMPA as /I, E, a, O, U, @/. About /a/ and /@/ (= IPA / ∂ /) we can agree. But the other symbols he proposes are inappropriately comparative. The Bulgarian vowels should appear in SAMPA as /i, e, a, o, u, @/.

B.1.3 Coding

SAMPA's coding principles involve restricting the available ASCII codes to the range 32–127. At the time SAMPA was formulated, many computers

used only the 7-bit ASCII character set. With the spread of PCs and compatibles, the "extended ASCII" (8-bit) set has become familiar, allowing codes in the range 128–255. Has the decision to restrict SAMPA to the range 32–127 proved wise? Or should we now relax it?

In the (American) English extended ASCII character set used by PCs running MS-DOS, the range 128–255 is used to provide for the screen and printer a number of accented alphabetic letters, currency symbols, graphic symbols, and Greek and mathematical symbols. Those that are not available on the keyboard can be accessed by entering their ASCII number on the keypad while pressing the Alt key. Unfortunately, from the point of view of non-English-speaking Europeans, this extended ASCII fails to provide all the accented Latin letters needed for such languages as Portuguese, Icelandic, Czech, Polish and Hungarian. To remedy this shortcoming, a number of different "code pages" are now available, each providing a different set of characters in the 128–255 range. In the USA and the UK most PCs use code page 437 (International English), in Western Europe 850 (Multilingual Latin I), and in much of Eastern Europe 852 (Slavic Latin II).

Applications running under the popular front-end Windows use yet another character set, one known as "enhanced ANSI". This is identical with the ASCII set for 33–127; for 128–255 it offers its own specific choice of accented alphabetic and other characters, with codes different from ASCII.

The consequence is that in PC-compatible computing the code numbers in the range 128–255 (the "extended" characters) may currently have several different interpretations. Conversely, a given character may be coded in several different ways.

Consider the IPA symbols $/\alpha$ / and $/\delta$ /, both needed for the phonetic transcription of English. For reasons that seemed valid at the time (cf. Wells (1987: 95)), SAMPA assigned the former the code 123, which now appears on all Latin-alphabet PC screens as "{"; the latter was coded 68, "D".

Both "æ" and "ð" are now available on-screen for PCs running Windows. While "æ" is an ASCII character, with the extended code 145 (for those using code page 437 or 850), "ð" is not. But both are in the enhanced ANSI set, with codes 230 and 240 respectively. (Hence under Windows they can be accessed, if not on the keyboard, by keying Alt+0230 and Alt+0240; "æ" can also be accessed as Alt+145.)

However, a PC using code page 852 (Slavic) will display code 145 as an upper-case L with acute accent (L), 230 as "Š", and 240 as "-". With code page 860 (Portugal), 145 is "À", 230 " μ " and 240 " \equiv ".

Recently a number of phonetic fonts have become available for use under Windows. These comprise only phonetic symbols (perhaps with a few punctuation signs). Unfortunately they disagree extensively on key assignment and coding. On my PC I now have three TrueType phonetic fonts provided by the Summer Institute of Linguistics and four others of whose origins, I regret to say, I am uncertain. These fonts agree with SAMPA (but not ANSI) in assigning " δ " to code 68/D; but for "æ" they assign codes and keystrokes 81/Q (SIL Doulos/Manuscript/Sophia IPA), 60/< (Times IPA New), 64/@ (Tech Phonetic), and 233 (IPA Roman 1, IPA Plus).

B.1.4 Further languages

A number of other EC languages have been examined in the light of the SAMPA recommendations, and a short summary of the possible solutions for their special features is given here. For more details, see J. Wells, "Computer-coded phonetic transcription", Journal of the International Phonetic Association 17, No. 2, pp. 94–114, and the SAM Definition Phase Final Report (ESPRIT project 1541), January 1988.

Most of the minority languages of Europe such as Basque, Breton, Catalan, and Frisian can be transcribed adequately at a phonemic level without the need to change the principles of the present recommendation. Irish and Scottish Gaelic, however, require a decision for coding the palatalised (or "slender") consonants and the "double" nasals and laterals. Scottish Gaelic also has a back unrounded vowel series which does not occur in other EC languages. Welsh requires a solution for the voiceless alveolar lateral, represented in the orthography as "N.

We should like now to explore whether it would be suitable to extend SAMPA for application to other languages, including Chinese, and if so how.

The question of Chinese has arisen because of the prospect of a wider collaboration on speech research between University College London and the Chinese Academy of Sciences.

Chinese already has what appears to be a satisfactory machine-readable phonetic notation in the form of Pinyin, the romanisation that has for some years been standard in the People's Republic (though not in Taiwan). Pinyin is an ingenious quasi-phonemic notation. It includes a number of unconventional digraphs, together with unconventional uses of individual Latin letters. Thus *sh*, *ch*, and *zh* represent retroflex/postalveolar consonants of a type that would normally be written in SAMPA as [S, tS, dZ]. Pinyin *x*, *q*, *j* represent a corresponding series of alveolopalatal consonants, IPA [c, tc, dz], for which SAMPA does not currently cater. Pinyin *c* represents [ts], *y* [j], and *ng* [ŋ]. The close front rounded vowel [y] is written *u* where there would be no confusion with [u], but \ddot{u} where this confusion might arise. (This last Pinyin character is not actually machine-readable in our sense.)

Continuing to use Pinyin for Chinese but SAMPA for other languages would mean that characters such as "x, j" would have different meanings in different languages ("x" = alveolopalatal fricative, or velar fricative; "j" =

alveolopalatal affricate, or palatal approximant). But this is perhaps no worse than the "comparative" differences already present in the interpretation of some symbols (see above). The Pinyin notation "i" already covers a remarkable range of allophonic possibilities (including an r-coloured back vowel in *shi* and a slightly fricative central vowel in *si*). Are Chinese speech technologists happy with this degree of phonemic abstraction?

Tone is shown in Pinyin (if indeed it is shown) by superscript accent marks, thus $m\bar{a}$, $m\acute{a}$, $m\acute{a}$, $m\acute{a}$. These are not machine-readable in the SAMPA sense. The corresponding SAMPA tone-marks would be /"ma, 'ma, 'ma, `ma/. However these SAMPA signs have not proved popular, and perhaps ought to be changed. For Chinese, we could perhaps consider instead the use of numerals, thus "ma1, ma2, ma3, ma4".

B.2 SAMPA: Present status

Tables B.1 to B.6 present the system agreed among the representatives of eight European countries engaged in European collaborative speech technology assessment research (SAM). It is currently being tested in the transcription and labelling of European multilingual databases.

	Table	B.1: Vowels	
I	Open-	Front-	Ro

IPA	SAMPA	ASCH	Open–	Front-	Rounded-	$\operatorname{Comment}$
			Close	Back	Unrounded	
а	a	97	open	front	unrounded	
a	А	65	open	back	unrounded	
æ	{	123	near-open	front	unrounded	English bad
в	6	54	near-open	$\operatorname{cent} \operatorname{ral}$	unrounded	German Butter
σ	Q	81	open	back	rounded	
Э	0	79	open-mid	back	rounded	
е	е	101	close-mid	front	unrounded	
ε	Е	69	open-mid	front	unrounded	
ə	Q	64	mid	$\operatorname{cent} \operatorname{ral}$	unrounded	Schwa
3	3	51	mid	$\operatorname{cent} \operatorname{ral}$	unrounded	
i	i	105	close	front	unrounded	
Ι	I	73	near-close	front	unrounded	lax
0	0	111	close-mid	back	rounded	
ø	2	50	close-mid	front	rounded	
œ	9	57	open-mid	front	rounded	
Œ	&	38	open	front	rounded	
u	u	117	close	back	rounded	
U	U	85	near-close	back	rounded	lax
ŧ	}	125	close	$\operatorname{cent} \operatorname{ral}$	rounded	
Λ	V	86	open-mid	back	unrounded	
у	У	121	close	front	rounded	
Υ	Y	89	near-close	front	rounded	lax

IPA	SAMPA	ASCII	Voicing	Place	Manner
b	b	98	voiced	bilabial	plosive
с	с	99	voiceless	palatal	plosive
ç	С	67	voiceless	palatal	fricative
d	d	100	voiced	dental/alveol.	plosive
ð	D	68	voiced	dental	fricative
f	f	102	voiceless	labiodental	fricative
g	g	103	voiced	velar	plosive
γ	G	71	voiced	velar	fricative
h	h	104	voiceless	glottal	fricative
j	j	106		palatal	approximant
k	k	107	voiceless	velar	plosive
1	1	108		dental/alveol.	lateral approxim.
λ	L	76		palatal	lateral approxim.
m	m	109		bilabial	\mathbf{nasal}
n	n	110		dental/alveol.	\mathbf{nasal}
ր	J	74		palatal	\mathbf{nasal}
ŋ	Ν	78		velar	\mathbf{nasal}
р	Р	112	voiceless	bilabial	plosive
r	r	114		alveolar	trill
в, в	R	82		uvular	trill/fricative
s	S	115	voiceless	alveolar	fricative
ſ	S	83	voiceless	$\operatorname{postalveolar}$	fricative
t	t	116	voiceless	dental/alveol.	plosive
θ	Т	84	voiceless	dental	fricative
v	v	118	voiced	labiodental	fricative
w	W	119		labial-velar	$\operatorname{approximant}$
x	x	120	voiceless	velar	fricative
Ч	H	72		labial-palat.	$\operatorname{approximant}$
z	z	122	voiced	alveolar	fricative
3	Z	90	voiced	postalveolar	fricative
3	?	63		stød, glottal stop	

Table B.2: Consonants

Table B.3: Two character symbols

IPA	SAMPA	ASCII	Comment
З	e~	69,126	Nasalised vowel
aı	aI	97,73	Diphthong
t∫	tS	116, 83	Affricate
d.	rd	$114,\!100$	Retroflex consonant
θ	u0	$117,\!48$	Mid-close rounded central vowel in Swedish
n	=n	61,110	Syllabic /n/
	E/	69,47	Indeterminacy (for $/e/ - /E/$ in French)

Table	B.4:	Boundary	and	prosodic	features
-------	------	----------	----------------------	----------	----------

IPA	SAMPA	ASCII	Comment
I	:	58	Length mark
papa	"papa	34	Primary stress (also in Accent I words in Norwegian and Swedish)
~papa	""papa	34,34	Accent II words in Norwegian and Swedish
papa	%papa	37	Secondary stress
⁻ papa	-papa	45	Level tone (if followed by a tone group boundary)
,papa	'papa	39	Rising tone
, papa	`papa	96	Falling tone
., papa	`'papa	96, 39	Fall-rise
, papa	'`papa	$39,\!96$	Rise-fall
	\$	36	Syllable boundary
		124	Tone group boundary
	-	45	Separator

Table B.5: Currently under discussion

SAMPA	ASCII	Comment
*	42	Conjunctor
	46, 46, 46	Silent pause
\	92	Phonetic case shift (e.g. \F might be used to signal
		a shift into French and \setminus would terminate the shift.

Table B.6: Currently used in French work

SAMPA	ASCII	Comment
§	21	Phonological phrase
#	35	Word boundary
##	$35,\!35$	Absence of liaison
+	43	Morpheme boundary

SAMPA computer readable phonetic alphabet 691

B.3 The phonemic notation of individual languages

This section provides a brief outline of the phonemic distinctions in the languages of the eight countries engaged in the initial phase of the SAM project. The SAMPA notation for individual languages presented here is an update of the SAM 1992 version.

Additionally, conventions are provided for the languages Spanish, Portuguese, Greek, Bulgarian, Estonian, Hungarian, Polish, and Romanian. Further information is available at:

http://www.phon.ucl.ac.uk/home/sampa/home

The SAMPA tables for each language follow below.
B.3.1 Bulgarian

VOWELS

The vowel system of Contemporary Standard Bulgarian comprises 6 phonemes, as follows.

Symbol	GLOSS	Transcription
i	peak	pik
e	$\operatorname{sunshine}$	pek
a	again	pak
Q	\mathbf{but}	p@k
0	\mathbf{under}	pOt
u	$\operatorname{crack}!$	puk

In unstressed positions there is considerable qualitative reduction of vowels.

Consonants

The consonant system comprises 21 "hard" (non-palatalised) phonemes, as follows.

Symbol	GLOSS	TRANSCRIPTION
Р	again	pak
Ъ	\mathbf{ball}	bal
t	there	tam
d	I give	dam
k	how	kak
g	gas	gas
ts	czar	tsar
dz	$_{ m clang}$	dz@n
tS	teak	tSam
dZ	glass	dZam
f	fall	fal
v	wall	val
S	alone	sam
z	dice	zar
S	shawl	Sal
Z	pity	Zal
x	bad luck	xal
m	$\operatorname{displacement}$	max
n	us	nas
1	polish	lak
r	once	ras

There are also 18 "soft" (palatalised or palatal) consonants, which are shown by the symbol "'" (ASCII 39) written after the consonant symbol. Before /i, e/ they do not contrast with the corresponding "hard" consonants, but elsewhere they do.

Symbol	GLOSS	TRANSCRIPTION
р'	sang	p'al
Ъ'	white	b'al
t'	them	t'ax
d'	apart	d'al
k'	$\operatorname{prophet}$	k'ar
g'	rose	g'ul
ts'	whole	ts'al
dz'	have bad luck	dz'an
f'	Fyut (name)	f'ut
v'	feeble	v'al
s'	sowed	s'al
z'	be badly off	z'an
х'	Hume (name)	x'um
m'	wineskin	m'ax
n'	mute	n'am
1'	poured	l'ax
r'	cut	r'as
j	ate	jal

B.3.2 Danish

Consonants

The plosives are p b t d k g:

Symbol	Word	TRANSCRIPTION
Р	\mathbf{pande}	"pan@
b	\mathbf{bande}	"ban@
t	tand	tan?
d	dan	dan?
k	\mathbf{kalde}	"kal@
g	galde	"gal@

The fricatives are **f** s:

f	finde	"fen@
S	stand	sdan?

The approximants are $v\ D\ j\ h:$

v	\mathbf{vinde}	"ven@
D	bide	"bi:D@
j	$_{\rm Jul}$	ju:?1
h	\mathbf{hest}	hEsd

The nasals are m n N:

m	mile	"mi:10
n	ny	ny?
Ν	$_{ m lunge}$	"loN@

The liquids are 1 R:

1	land	lan?
R	ride	"Ri:D@

Stød is symbolised by ? and may be found in syllables containing a long stressed vowel, or a short stressed vowel, or a short stressed vowel followed by a voiced consonant, e.g. $p{u: -pE:?u, peu: -pEu?}$

PURE VOWELS

The vowel system chosen for broad phonetic transcription is $i \in E = A y 2 9 u \circ 0 0$, with all vowels except 0 occurring with a length distinction:

 $(i: e: E: \{: A: y: 2: 9: u: o: 0:).$

The unrounded front vowels are exemplified in the following:

Symbol	Word	Transcription
i:	$_{ m mile}$	"mi:10
i	ville	"vil@
e:	\mathbf{mele}	"me:10
e	\mathbf{visse}	"ves@
E :	mæle	"mE:10
Е	$\mathrm{t}\mathbf{\hat{x}}\mathrm{k}\mathrm{k}\mathrm{e}$	"tEk@
{:	\mathbf{male}	"m{:10
a	\mathbf{malle}	"mal@

The central vowels are:

Symbol	Word	Transcription	
A :	parken	"pA:g=n	
Α	\mathbf{pakken}	pAg=n	
Q			unstressed final e

The rounded front vowels are:

у:	hy le	"hy:l@
у	$_{ m hylde}$	"hyl@
2:	køle	"k2:10
2	kølle	"k210
9:	høne	"h9:n@
9	hønse	"h9ns@

The back vowels are:

u:	kule	"ku:10
u	kulde	"kul@
o:	fole	"fo:l@
0	foto	"fodo
0:	${ m m}{ m ane}$	"mO:n@
0	munde	"mOn@
Q:	kåre	"kQ:@
Q	kors	kQs

DIPHTHONGS.

The falling diphthongs may be most economically analysed phonemically as vowel plus j, v, or r, but for broad phonetic representation within SAMPA they are analysed as vowel plus i, u or Q, for example:

Symbol	Word	TRANSCRIPTION	
Ai	sejle	"sAil@	
Qi	søjle	"sQil0	
ui	huje	"hui@	extremely rare
i:i	pige	"pi:i@	
e:i	ege	"e:i@	conservative pronunciation
E:i	læge	"lE:i@	
iu	ivrig	"iuRi	
eu	peber	"peu@R	
Eu	evne	"Eun@	
i:u	hive	"hi:u	
e:u	leve	"le:u	
E:u	hæve	"hE:u	
au	havn	hau?n	
yu	syv	syu	
2u	$d \phi vstum$	"d2usdom	
9u	støvle	"sd9ul@	
iQ	kirke	"kiQk@	
EQ	bær	bEQ	
9Q	sm ør	sm9Q	

B.3.3 Dutch

Consonants

The plosives are p b t d k (g):

Symbol	Word	Transcription	
Р	$_{ m pak}$	pAk	
b	\mathbf{bak}	bAk	
t	tak	tAk	
d	dak	dAk	
k	$_{\mathrm{kap}}$	kAp	
g	goal	go:l	only foreign/loanwords

The fricatives are f v s z x (G) h:

f	fel	fEl
v	vel	vEl
S	sein	sEin
z	zijn	zEin
x	toch	tOx
G	goed	Gut (also: xut)
h	\mathbf{hand}	hAnt
Z	bagage	bAga:Z(@)
S	show	So:u
Z S	bagage show	bAga:Z(@) So:u

The sonorants (nasals, liquids and glides) are m n N l r w j:

m	met	mEt
n	\mathbf{net}	nEt
Ν	\mathbf{bang}	bAN
1	land	lAnt
r	rand	rAnt
W	wit	wIt
j	ja	ja:

VOWELS

The Dutch vowels fall into two classes, "checked" (not occurring in a stressed syllable without a following consonant) and "free".

The checked vowels are I $\, E\, \, A\,\, O\,\, Y\,\, {\tt Q}:$

Symbol	Word	TRANSCRIPTION
I	pit	pIt
Е	pet	pEt
А	pat	pAt
0	pot	pOt
Y	put	pYt
Q	$\operatorname{gemakkelijk}$	G@"mAk@l@k

The free vowels comprise four monophthongs i y u a:, three "potential diphthongs" e: 2: o:, and three "essential diphthongs" Ei 9y Au, exemplified as follows:

Symbol	Word	TRANSCRIPTION
i	vier	vir
У	vuur	vyr
u	voer	vur
a:	naam	na:m
e:	veer	ve:r
2:	deur	d2:r
o:	voor	vo:r
Ei	fijn	fEin
9y	huis	h9ys
Au	goud	xAut

There are also six vowel sequences which are sometimes described as diphthongs:

a:i	draai	"dra:i
o:i	mooi	"mo:i
ui	$\operatorname{roeiboot}$	"ruibo:t
iu	nieuw	"niu
yu	duw	"dyu
e:u	sneeuw	"sne:u

Several marginal vowel phonemes are only found in loanwords:

Word	TRANSCRIPTION
$\operatorname{cr}\operatorname{e}\mathbf{m}\mathrm{e}$	krE:m
freule	fr9:10
roze	r0:z@
	WORD crème freule roze

B.3.4 English

Consonants

The standard English consonant system is traditionally considered to comprise 17 obstruents (6 plosives, 2 affricates and 9 fricatives) and 7 sonorants (3 nasals, 2 liquids and 2 semivowel glides).

With the exception of the fricative /h/, the obstruents are usually classified in pairs as "voiceless" and "voiced", although the presence or absence of periodicity in the signal resulting from laryngeal vibration is not a reliable feature distinguishing the two classes. They are better considered "fortis" (strong) and "lenis" (weak), with duration of constriction and intensity of the noise component signalling the distinction.

The six plosives are p b t d k g:

Symbol	Word	TRANSCRIPTION
Р	pin	pIn
b	bin	bIn
t	$_{ m tin}$	tIn
d	din	dIn
k	$_{ m kin}$	kIn
g	give	gIv

The "lenis" stops are most reliably voiced intervocalically; aspiration duration following the release in the fortis stops varies considerably with context, being practically absent following /s/, and varying with degree of stress syllable-initially.

The two phonemic affricates are tS and dZ:

Symbol	Word	Transcription
tS	$_{ m chin}$	tSIn
dZ	$_{ m gin}$	dZIn

As with the lenis stop consonants, /dZ/ is most reliably voiced between vowels.

There are nine fricatives f v T D s z S Z h:

700 Part IV: Spoken language reference materials

Symbol	Word	TRANSCRIPTION
f	fin	fIn
v	vim	vIm
Т	thin	TIn
D	this	DIs
S	\sin	sIn
z	zing	zIN
S	$_{ m shin}$	SIn
Z	${ m measure}$	"meZ@
h	hit	hIt

Intervocalically the lenis fricatives are usually fully voiced, and they are often weakened to approximants (fricationless continuants) in unstressed position.

The sonorants are three masals m n N, two liquids r 1, and two sonorant glides w j:

Symbol	Word	Transcription
m	mock	mQk
n	knock	nQk
N	thing	TIN
r	wrong	rQN
1	\log	lQN
W	wasp	wQsp
j	yacht	jQt

Vowels

The English vowels fall into two classes, traditionally known as "short" and "long" but, owing to the contextual effect on duration of following "fortis" and "lenis" consonants (traditional "long" vowels preceding fortis consonants can be shorter than "short" vowels preceding lenis consonants), they are better described as "checked" (not occurring in a stressed syllable without a following consonant) and "free".

The checked vowels are I e { Q V U:

Symbol	Word	TRANSCRIPTION
I	pit	pIt
е	pet	pet
{	pat	p{t
Q	pot	pQt
V	cut	kVt
U	put	pUt

There is a short central vowel, normally unstressed:

Q another Q"nVDQ

The free vowels comprise monophthongs and diphthongs, although no hard and fast line can be drawn between these categories. They can be placed in three groups according to their final quality: i: eI aI OI, u: @U aU, 3: A: O: I@ e@ U@. They are exemplified as follows:

Symbol	Word	Transcription
i:	ease	i:z
eI	\mathbf{raise}	reIz
aI	rise	raIz
OI	\mathbf{noise}	nOIz
u:	lose	lu:z
QU	nose	n@Uz
aU	rouse	raUz
3:	\mathbf{furs}	f3:z
A :	stars	stA:z
0:	cause	k0:z
IQ	fears	fI@z
e@	stairs	ste@z
U@	cures	kjU@z

The vowels /i:/ and /u:/ in unstressed syllables vary in their pronunciation between a close [i]/[u] and a more open [I]/[U]. Therefore, it is suggested that /i/ and /u/ be used as indeterminacy symbols:

i	happy	"h{pi
u	into	"Intu

- 1. NOTATIONAL VARIANTS. Differently from the notation set out above:
 - It is possible to transcribe English long vowels without using length marks, thus /i u 3 A 0/. This is phonemically unambiguous, although it does remove the option of restricting the symbols [i u] to the use just described, for the phonemically indeterminate weak vowels.
 - 2. The symbol /E/ is quite widely used in place of /e/ for the vowel of "pet".
 - 3. In an older notation, now no longer in general use, paired short and long vowels were transcribed using the same vowel symbol with and without length marks, thus /i/ in "pit", /i:/ in "ease"; /0/ in "pot", /0:/ in "cause".

- 702 Part IV: Spoken language reference materials
- 2. ADDITIONAL SYMBOLS. For some purposes and some varieties of English it is useful to give explicit symbolisation to the glottal stop and/or the voiceless velar fricative:

?	network	ne?w3:k
x	loch	lQx

B.3.5 Estonian

VOWELS

The standard Estonian vowel system comprises 9 vowels, all of which can occur single (short) or doubled (long). The vowel /7/ is high-mid back unrounded, in the area of cardinal 15.

Symbol	Word	TRANSCRIPTION
i	risu	risu
ii	riisu	riisu
e	kere	kere
ee	keere	keere
{	käru	k{ru
{{	kääru	k{{ru
У	nüri	nyri
уу	müüri	myyri
2	köhi	k2hi
22	nööri	n22ri
u	kuri	kuri
uu	kuuri	kuuri
0	pori	pori
00	poori	poori
7	sõre	s7re
77	sõõre	s77re
A	varu	vAru
AA	vaaru	vAAru

Consonants

The consonant system comprises 17 consonant phonemes, all of which can occur single (short) or double (long). Those symbolised with /'/ are palatalised.

Symbol	Word	Transcription
Р	taba	tApA
рр	tapa	tAppA
t	padu	pAtu
tt	patu	pAttu
k	kagu	kAku
kk	kaku	kAkku
t'	padi	pAt'i
t't	pati	pAt'ti
f	foori	foori
ff	tufi	tuffi
v	tuvi	tuvi
vv	kavva	k A v v A

Symbol	Word	TRANSCRIPTION
S	kasu	kAsu
SS	kassa	kAssA
S	s <efi looz<i<="" td=""><td>Seffi looSi</td></efi>	Seffi looSi
SS	tus < i	tuSSi
h	sahin	sAhin
hh	m s <ahhi< td=""><td>SAhhi</td></ahhi<>	SAhhi
s'	kasi	kAs'i
s's	kassi	kAs'si
m	samas	sAmAs
mm	sammas	s AmmA s
n	kana	k An A
nn	kanna	k Ann A
n'	pani	pAn'i
n'n	panni	pAn'ni
1	$_{\rm kalas}$	kAlAs
11	kallas	kAllAs
1'	pali	pAl'i
1'1	palli	pAl'li
r	varas	vArAs
rr	varras	vArrAs
j	maja	mAjA
jj	majja	mAj:jA

704 Part IV: Spoken language reference materials

WORD ACCENT

There are two types of word accent: "light" and "heavy". The heavy accent implies additional duration, and is shown in SAMPA by using the length mark after the appropriate segmental symbol. The interaction of these two accents with single and double segments produces segmental contrasts of short vs. long vs. overlong.

(short, light)	nagi	nAki
(long cons., light)	$_{ m maki}$	mAkki
(long cons., heavy)	$_{ m makki}$	mAk:ki
(long vowel, light)	${ m maagi}$	mAAki
(long vowel, heavy)	${ m maagi}$	mAA:ki

B.3.6 French

 $\operatorname{Consonants}$

The standard consonant system is considered to consist of 12 obstruents (six plosives and six fricatives) and 8 sonorants (three nasals, two liquids, and three semivowel glides). The obstruents can be classified in voiced and voiceless pairs, with strong periodicity (voicing) normally occurring in the phonemically voiced members.

The plosives are p b t d k g:

Symbol	Word	Transcription
Р	pont	po~
Ъ	bon	bo~
t	$_{ m temps}$	ta~
d	$_{ m dans}$	da~
k	quand	ka~
g	gant	ga~

The voiceless plosives (/p t k/) are unaspirated except in stressed syllables preceding close vowels, where the extreme position of the tongue delays voice onset and produces turbulence.

There are six fricatives f v s z S Z; there is also j, which may be considered a fricative or a glide:

Symbol	Word	Transcription
f	femme	fam
v	vent	va~
S	sans	sa~
z	zone	zon
S	champ	Sa~
Z	$_{\rm gens}$	Za~
j	ion	jo~

There are three nasals, m n J, found in words considered to be genuinely French. A fourth nasal, N, is only found in loanwords, except in Southern French dialects, where it occurs in some contexts after nasal vowels:

Symbol	Word	TRANSCRIPTION
m	mont	mo~
n	nom	no~
J	oignon	o Jo~
Ν	$\operatorname{camping}$	ka~piN

There are two liquids, 1 R, and three vowel glides, w H and j. The vowel glides may be realised as fricative following voiceless obstruents.

Symbol	Word	Transcription
1	\log	lo~
R	rond	Ro~
W	coin	kwe~
Н	juin	ZHe~
j	pierre	pjER

VOWELS

The vowel system comprises 12 oral vowels, i e E = A O o u y 2 9 O, and 4 nasal vowels, $e^{-} a^{-} o^{-} 9^{-}$, exemplified as follows:

Symbol	Word	Transcription
i	si	si
е	ses	se
Е	seize	sEz
a	\mathbf{patte}	pat
A	\mathbf{p} âte	pAt
0	comme	kOm
0	gros	gRo
u	doux	du
У	du	dy
2	deux	d2
9	\mathbf{neuf}	n9f
Q	${ m justement}$	$\texttt{Zyst@ma}^{\sim}$
e~	vin	ve~
a~	\mathbf{vent}	va~
o~	bon	bo~
9~	brun	bR9~

When they are functional, the load of the oppositions a-A, $e^{-9^{\circ}}$, e-E, o-0, 2-9 may be very low for certain speakers, and there is a tendency towards neutralisation. When they are not functional there is a strong tendency in unstressed syllables towards indetermination. "Indeterminacy" symbols have been agreed to cover occurrences of these phonemes or sounds.

E/=e or E O/=o or O &/ = 2 or 9 A/=a or A $U^{\sim}=e^{\sim} \text{ or } 9^{\sim}$

There are contextually determined vowel length differences, nasal vowels being long before following consonants, and all vowels being long before R and voiced fricatives.

B.3.7 German

CONSONANTS

The standard German consonant system is considered to have 17 or 19 obstruent phonemes (depending on whether two peripheral sounds are included, which occur only in loanwords), and five sonorants. The obstruents comprise six plosives, three (or four) affricates, and eight (or nine) fricatives, though there are two auditorily distinct fricatives (x and C) which are usually considered to be allophonic variants, giving ten fricatives in all that require distinct symbolic representation. [For some purposes it is convenient to give explicit representation to the glottal stop, too.]

As in English, the obstruents are traditionally classified pairwise as "voiced" and "voiceless", though periodicity is a less reliable feature than duration and intensity, and they are therefore better termed "lenis" and "fortis".

The six plosives are p b t d k g:

pPeinpaInbBeinbaIntTeichtaICdDeichdaICkKunstkUnstgGunstgUnst	Symbol	Word	Transcription
bBeinbaIntTeichtaICdDeichdaICkKunstkUnstgGunstgUnst	Р	Pein	paIn
t Teich taIC d Deich daIC k Kunst kUnst g Gunst gUnst	Ъ	Bein	baIn
d Deich daIC k Kunst kUnst g Gunst gUnst	t	Teich	taIC
k Kunst kUnst g Gunst gUnst	d	Deich	daIC
g Gunst gUnst	k	Kunst	kUnst
	g	Gunst	gUnst

The voicing and aspiration patterning of German plosives is similar to that of English, with the unaspirated variant occurring after initial /S/ or (in a few words) /s/.

[If it is desired to symbolise the glottal stop explicitly, it may be shown in SAMPA as shown here.]

? Verein fE6"?aIn

There are three phonemic affricates that are considered to be native to German, pf ts tS, and a fourth one, dZ, which occurs in a few loanwords, and which is often replaced by tS:

Symbol	Word	Transcription
pf	$\mathbf{P}\mathbf{f}\mathbf{a}\mathbf{h}\mathbf{l}$	pfa:l
ts	Zahl	tsa:1
tS	$\operatorname{deutsch}$	dOYtS
dZ	Dschungel	"dZUN=1

There are ten fricatives, f v s z S Z C j x h. j is often realised as a

vowel glide:

Symbol	Word	Transcription
f	fast	fast
v	was	vas
S	Tasse	"tas@
z	Hase	"ha:z@
S	waschen	"vaS=n
Z	Genie	Ze"ni:
С	sicher	"zIC6
j	$_{ m Jahr}$	ja:6
x	Buch	bu:x
h	Hand	hant

The sonorants are three nasals, m n N, and two "liquids", 1 R, of which R can be realised as a uvular fricative (voiced or voiceless depending on context), a uvular approximant, or a uvular tap or trill:

Symbol	Word	Transcription
m	mein	maIn
n	nein	naIn
Ν	Ding	dIN
1	Leim	laIm
R	Reim	RaIm

Orthographic <r> is realised phonetically in a number of different ways:

- 1. As a dorso-uvular consonant a voiced or voiceless fricative, approximant, trill or flap. This should be represented as R (as above).
- 2. As an apico-alveolar consonant a trill, tap, or flap. This may be represented as r, e.g. <Reim> raIm.
- 3. As a vowel post-vocalically. This may be represented as 6 (see below).

VOWELS

The vowels fall into three groups, "checked" (short), "free" (long), and two short vowels that only occur in unstressed position. There is a genuine short-long vowel distinction in German, the long vowels being roughly twice as long (all other things being equal) as the short vowels.

The checked vowels are I $\tt E$ a O U Y 9:

Symbol	Word	TRANSCRIPTION
I	Sitz	zIts
Е	Gesetz	g@"zEts
a	Satz	zats
0	Trotz	trOts
U	$\operatorname{Schut} z$	SUts
Y	${ m h\ddot{u}bsch}$	hYpS
9	$\operatorname{pl\"ot} \operatorname{zlich}$	"pl9tslIC

There are 8 pure free vowels, i: e: E: a: o: u: y: 2:, and three free diphthongs, aI aU OY:

Symbol	Word	Transcription
i:	Lied	li:t
e:	Beet	be:t
Е:	spät	SpE:t
a:	Tat	ta:t
o:	rot	ro:t
u:	Blut	blu:t
у:	süß	zy:s
2:	blöd	b12:t
aI	Eis	als
aU	Haus	haUs
OY	Kreuz	krOYts

The unstressed "schwa" vowel is:

The vowel realisation of <r>, represented as 6, fuses with schwa, but it also follows stressed vowels, resulting in additional centring diphthongs:

Symbol	Word	TRANSCRIPTION
6	$_{\mathrm{besser}}$	"bEs6
i:6	Tier	ti:6
16	Wirt	vI6t
у:6	Tür	ty:6
¥6	Türke	"tY6k@
e:6	schwer	Sve:6
E6	Berg	bE6k
E:6	Bär	bE:6
2:6	Föhr	f2:6
96	Wörter	"v96t6
a:6	Haar	ha:6
a6	hart	ha6t
u:6	Kur	ku:6
U6	kurz	kU6ts
o:6	Ohr	o:6
06	dort	d06t

$B.3.8 \quad Greek$

CONSONANTS:

Plosives:

Symbol	Word	Transcription
Р	$\pi i u \omega$	"pino
Ъ	$\mu\pi oarrho\omega$	bo"ro
t	$\tau \acute{o} \tau \varepsilon$	"tote
d	$\nu \tau \dot{\upsilon} \nu \omega$	"dino
k	$\kappa \dot{lpha} u \omega$	"kano
g	$\gamma\kappaarrhoarepsilon\mu i\zeta\omega$	gre"mizo

Affricates:

ts	$ au \sigma arepsilon \pi \eta$	"tsepi
dz	$ au\zetalpha\mu\pilpha au\zeta\etaarsigma$	dzaba"dzis

Fricatives:

Symbol f v	$egin{array}{c} WORD \ arphi arepsilon \psi arepsilon \psi arphi \omega \ eta arphi arphi arphi \omega \ eta arphi ar$	TRANSCRIPTION "fevGo "vraDi	
	θελω	"lelo	
D	ο ορομος	"Dromos	
S	$\sigma \pi i au \iota$	"spiti	
z	$\zeta arepsilon \sigma au \eta$	"zesti	
x	χρόνος	"xronos	
G	$\gamma \dot{lpha} \lambda lpha$	"Gala	
Nasals:			
Symbol	Word	Transcription	
m	$\mu\iota\lambda\dot{\omega}$	mi"lo	
n	νερό	ne"ro	
	αγγούρι	an"guri	$[\rm N]$ = all ophone of /n/
Liquids:			

Symbol	Word	TRANSCRIPTION
1	$\lambda o v \lambda o \acute{v} \delta \iota$	lu"luDi
r	$arrho v \chi lpha$	"ruxa

Semivowel:

i	διαβάζω	Dia"wazo
J	υπαραζω	שן א מצט

SAMPA computer readable phonetic alphabet 711

(Palatals)

καί	ce
κοιμάμαι	ci"mame
σ κιάζ ω	"scazo
$lpha\gamma\gammaarepsilon\lambda oarsigma$	"angjelos
$\gamma\kappa i u \imath lpha$	"gjinja
$\gamma\kappa\iota ovetaarepsilon au\sigma\iota$	gju"vetsi
χ έ $ ho$ ι	"Ceri
$\chi arepsilon \iota \mu \dot{\omega} u lpha arsigma$	Ci"monas
<i>χιόνι</i>	"Coni
$\gamma \acute{e} \varrho o \varsigma$	"jjeros
$\gamma \acute{v} \varrho o \varsigma$	"jjiros
γ ιός/ v ιός	jjos
	καί κοιμάμαι σκιάζω άγγελος γκίνια γκιουβέτσι χέρι χειμώνας χιόνι γέρος γύρος γιός/υιός

Connected speech phenomena:

δέν πειράζει	(Den pi"razi)	De bi"razi
ο γιός μου	(o "jjos mu)	o "jjoz mu

B.3.9 Hungarian

VOWELS

The vowel system comprises 14 phonemes, as follows:

Symbol	Word	TRANSCRIPTION
i	hit	hit
i:	szít	si:t
Е	vet	vEt
e:	méz	me:z
0	hat	hOt
A :	láb	la:b
0	sok	sok
o:	pók	po:k
2	köt	k2t
2:	sőt	S2:t
u	fut	fut
u:	kút	ku:t
У	süt	Syt
у:	fűt	fy:t

Consonants

The consonant system comprises 24 phonemes, as follows:

Symbol	Word	TRANSCRIPTION
Р	Pál	pa:l
b	bál	ba:l
t	tár	ta:r
d	dán	da:n
t'	tyúk	t'u:k
d'	gyár	d'a:r
k	kád	ka:d
g	gát	ga:t
ts	cél	tse:1
dz	bodza	bodza
tS	cső	tS2:
dZ	dzsem	dZEm

SAMPA computer readable phonetic alphabet 713

Symbol	Word	TRANSCRIPTION
f	fát	fa:t
v	vád	va:d
S	szép	se:p
z	zár	za:r
S	só	So:
Z	zsír	Zi:r
m	már	ma:r
n	\mathbf{n} ád	na:d
J	nyom	Jom
r	rák	ra:k
1	láp	la:p
j	jön, lyuk	j2n, juk
h	hát	ha:t
x	achát	Oxa:t

It is convenient for some purposes to symbolise separately the following salient allophones:

F	kámfor	ka:Ffor
N	ing	iNg

Consonant length is distinctive in Hungarian. All consonants may be phonemically short or long (geminate); this contrast operates in intervocalic and final-postvocalic positions.

B.3.10 Italian

Consonants

The Italian consonant system has 29 obstruents and 14 sonorants. The obstruents are classified along two dimensions, voiced vs. voiceless and single vs. geminate; the sonorants are (by definition) all voiced, but are also distinguished on the single-geminate axis, with the exception of two semivowel glides.

The voiced vs. voiceless distinction is normally manifested acoustically by the presence vs. absence of periodicity resulting from vocal fold vibration (cf. French).

There are six single and six geminate plosives, p b t d k g, pp bb tt dd kk gg, as follows:

Symbol	Word	TRANSCRIPTION
Р	\mathbf{pane}	"pane
Ъ	$_{ m banco}$	"banko
t	tana	"tana
d	danno	"danno
k	cane	"kane
g	$_{ m gamba}$	"gamba
РР	coppa	"kOppa
bb	$_{ m gobba}$	"gObba
tt	zitto	"tsitto
dd	cadde	"kadde
kk	nocca	"nOkka
gg	${ m fugga}$	"fugga

The plosives t, tt, d, dd have a dental, not an alveolar place of articulation.

There are four single and four geminate affricates, ts dz tS dZ, tts ddz ttS ddZ:

Symbol	Word	TRANSCRIPTION
ts	zitto	"tsitto
dz	zona	"dzOna
tS	cena	"tSena
dZ	gita	"dZita
tts	bozza	"bOttsa
ddz	mezzo	"mEddzo
ttS	braccio	"brattSo
ddZ	oggi	"OddZi

There are five single and four geminate fricatives, f v s z S, ff vv ss SS:

Symbol	Word	TRANSCRIPTION	
f	fame	"fame	
v	vano	"vano	
S	sano	"sano	
z	$\operatorname{sbaglio}$	"zbaLLo	initial only in cluster
S	scendo	"Sendo	
ff	beffa	"bEffa	
vv	bevvi	"bevvi	
SS	cassa	"kassa	
SS	ascia	"aSSa	

There are three single and geminate nasals, m n J, mm nn JJ, three single and three geminate liquids, r l L, rr ll LL, and two semivowels, j w:

Symbol	Word	Transcription	
m	molla	"mOlla	
n	nocca	"nOkka	
J	gnocco	"JOkko	
mm	grammo	"grammo	
nn	panna	"panna	
JJ	bagno	"baJJo	
r	rete	"rete	
1	lama	"lama	
L	gli	Li	only in function words
rr	ferro	"fErro	
11	colla	"kOlla	
LL	foglia	"fOLLa	
j	ieri	"jEri	
W	uomo	''w0mo	

Note: The velar nasal N is an allophone of n used before a velar. It may be written either way, thus < banco > banko or baNko.

There is no single-geminate opposition for the following six consonants in intervocalic position (including across word boundaries):

Five are always geminate: tts ddz SS JJ LL One is always single: z

VOWELS								
The vowel system	comprises seven	vowels, i	е	Е	a	0	0	u

Symbol	Word	TRANSCRIPTION
i	mite	"mite
е	rete	"rete
Е	meta	"mEta
a	rata	"rata
0	moto	"mOto
0	dove	"dove
u	muto	"muto

In unstressed position, the oppositions ${\tt e}$ – ${\tt E}$ and ${\tt o}$ – ${\tt O}$ are often neutralised.

B.3.11 Norwegian

Consonants There are six plosives:

Symbol	Word	TRANSCRIPTION
Р	hopp	hOp
Ъ	labb	lAb
t	lat	lA:t
d	ladd	lAd
k	takk	tAk
g	tagg	tAg

There are six fricatives:

Symbol	Word	TRANSCRIPTION	
f	$_{ m fin}$	fi:n	
v	vin	vi:n	
S	lass	lAs	
S	$_{ m skyt}$	Sy:t	
С	kino	Ci:nu	not syllable-final
j	gi	ji:	
h	ha	hA:	

There are five sonorant consonants (nasals, liquids, trills):

m	$_{ m lam}$	lAm
n	vann	vAn
N	sang	s AN
1	$_{\mathrm{fall}}$	fAl
r	prøv	pr2:v

Vowels

There are 9 long vowels:

i:	vin	vi:n
e:	sen	se:n
{:	vær	v{:r
A :	hat	hA:t
у:	lyn	ly:n
2:	søt	s2:t
0:	våt	v0:t
u:	bok	bu:k
}:	lun	1}:n

and nine short vowels:

i	vind	vin
е	send	sen
{	vært	v{rt
A	hatt	hAt
У	lynne	lyne
2	søtt	s2t
0	vått	vOt
u	bukk	buk
}	lund	1}n

There are seven diphthongs:

{i	vei	v{i
2у	høy	h2y
A }	sau	sA}
Ai	$_{\rm kai}$	kAi
0 у	konvoy	kun v0 y
}i	hui	h}i
ui	hoi	hui

In addition there are important allophonic variants for which the transcription has been agreed:

\mathbf{Symbol}	Word	Transcription	
rt	hardt	hArt	retroflex t
rd	verdi	v{rdi:	retroflex d
rl	pproxrlig	{:rlig	retroflex l
rn	$_{\mathrm{garn}}$	gA:rn	retroflex n
rL	blå	brLO:	retroflex flap

In cases where the dental consonants do not change into retroflexes, they are transcribed using the separator sign (ASCII 45), e.g.:

r-d verdig v{:r-di

B.3.12 Polish

VOWELS

The vowel system comprises 8 phonemes ("~" means nasalisation):

Symbol	Word	TRANSCRIPTION
i	PIT	pit
I	$_{ m typ}$	tIp
е	test	test
a	pat	pat
0	pot	pot
u	puk	puk
e~	gęś	ge~s'
o~	wąs	vo~s

Consonants

The consonant system comprises 29 phonemes (' means palatalisation):

Symbol	Word	Transcription
Р	pik	pik
Ъ	bit	bit
t	test	test
d	$_{ m dym}$	dIm
k	kit	kit
g	gen	gen
f	$_{ m fan}$	fan
v	wilk	vilk
S	syk	sIk
z	zbir	zbir
S	szyk	SIk
Z	żyto	ZIto
s'	świt	s'vit
z'	źle	z'le
x	$_{ m hymn}$	x Imn
ts	cyk	tsIk
dz	dzwon	dzvon
tS	czyn	tSIn
dZ	$d\dot{z}em$	dZem
ts'	$\acute{\mathrm{cma}}$	ts'ma
dz'	dźwig	dz'vik
m	mysz	mIS
n	nasz	naS
n'	koń	kon'
Ν	\mathbf{pek}	peNk
1	luk	luk
r	ryk	rIk
W	łyk	wIk
j	jak	jak

B.3.13 Portuguese

Consonants

Plosives:

Symbol	Word	TRANSCRIPTION
Р	pai	pai
Ъ	barco	"barku
t	tenho	"teJu
d	doce	"dos@
k	com	ko~
g	grande	"gr6nd@

Fricatives:

f	$_{\mathrm{falo}}$	"falu
v	\mathbf{verde}	"verd@
s	céu	sEw
z	casa	"kaz6
S	$\operatorname{chap\acute{e}u}$	S6"pEw
Z	jóia	"ZOj6

Nasals:

m	mar	mar
n	\mathbf{nada}	"nad6
J	$_{\rm vinho}$	"viJu

Liquids:

1	lanche	"16nS@
L	${ m trabalho}$	tr6"baLu
r	caro	"karu
R	rua	"Ru6

VOWELS AND DIPHTHONGS

Symbol	Word	TRANSCRIPTION
i	vinte	"vint@
	lápis	"lapiS
е	fazer	f6"zer
Е	belo	"bElu
a	falo	"falu
6	cama	"k6m6
	$\operatorname{madeira}$	m6"d6jr6

SAMPA computer readable phonetic alphabet 721

Symbol O o u C	WORD ontem lobo jus futuro felizes	TRANSCRIPTION "Ont6~j~ "lobu ZuS fu"turu f@"liz@S	
i~ e~ 6~ u~	fim emprego irmã bom um	fi [~] e [~] "pregu ir"m6 bo [~] u [~]	(or em-)
aw aj 6~j~	mau mais têm	maw majS t6~j~	etc.: iw, ew, Ew, (ow) etc.: ej, Ej, Oj, oj, etc.: e~j~, o~j~, u~j~

B.3.14 Romanian

VOWELS

The vowel system comprises 8 phonemes, as follows. The vowel symbolised i_0 is voiceless; 1 is a close (high) unrounded central vowel.

Symbol	Word	Transcription
i	vin	vin
i_0	$c\hat{a}$ ini	k1jni_0
е	$_{\mathrm{fel}}$	fel
a	cap	kap
Q	${f m}{f a}{f r}$	m@r
0	loc	lok
u	sur	sur
1	$_{ m fan}$	f1n
ע ורד	• 1	

There are 4 semivowels:

Symbol	Word	Transcription
j	doi	doj
e_X	${ m deas}$ ă	de_Xas@
W	sau	saw
0_X_0	$\operatorname{culoare}$	kulo_Xare

Consonants

The consonant system comprises 20 phonemes, as follows:

Symbol	Word	Transcription
Р	$\mathbf{p}\mathbf{\check{a}}\mathbf{r}$	p@r
b	barbă	barb@
t	tun	tun
d	dar	dar
k	cal	kal
g	$g \hat{a} t$	g1t
ts	ţară	tsar@
tS	ceas	tSas
dZ	$\operatorname{geant \breve{a}}$	dZant@
f	fată	fat@
v	vale	vale
S	scaun	skaun
z	zi	zi
S	şa	Sa
Z	joi	Zoj
h	${f hain}{f a}$	hajn@
m	${f m} {f a} {f n} {f a}$	m1n@
n	\mathbf{nas}	nas
1	lapte	lapte
r	$r\hat{a}s$	r1s

B.3.15 Spanish

Consonants

Plosives:

Symbol	Word	TRANSCRIPTION
Р	padre	"paDre
Ъ	vino	"bino
t	tomo	"tomo
d	donde	"donde
k	casa	"kasa
g	gata	"gata

Affricates:

Symbol	Word	TRANSCRIPTION
tS	mucho	"mutSo
jj	\mathbf{hielo}	"jjelo

Fricatives:

f B T D s x	fácil cabra cinco nada sala mujer	"faTil "kaBra "Tinko "naDa "sala mu"xer	= /b/ = /d/
x G	luego	"lweGo	= /g/

Nasals:

m	mismo	"mismo
n	nunca	"nunka
J	año	"aJo

Liquids:

1	lejos	"lexos	
L	$\operatorname{caballo}$	ka"baLo	or as jj
r	puro	"puro	
rr	torre	"torre	

Semivowels:

j	rei	rrej
	pie	pje
W	deuda	"dewDa
	muy	mwi
Vowels		
i	pico	"piko
е	pero	"pero
a	valle	"baLe
0	toro	"toro
u	duro	"duro

B.3.16 Swedish

CONSONANTS There are six plosives:

Symbol	Word	TRANSCRIPTION
Р	pil	pi:l
b	bil	bi:l
t	tal	tA:1
d	dal	dA:1
k	kal	kA:1
g	gås	go:s

There are six fricatives:

f	fil	fi:1	
v	vår	vo:r	
s	sil	si:l	
S	$_{ m sjuk}$	S } : k	front and back allophones
h	hal	hA:1	
С	t jock	COk	not syllable-final

There are six sonorant consonants (nasals, liquids and semi-vowels):

m	mil	mi: l	
n	\mathbf{n} ål	no: l	
N	ring	rIN	not syllable-initial
r	\mathbf{ris}	ri:s	
1	lös	12:s	
j	$_{ m jag}$	jA:g	

VOWELS

There are nine long and nine short vowels:

Long vowels (followed by a short consonant):

Symbol	Word	TRANSCRIPTION
i:	vit	vi:t
e:	vet	ve:t
E :	säl	sE:1
у:	syl	sy:1
}:	hus	h
2:	föl	f2:1
u:	sol	su:1
o:	hål	ho:l
A :	hal	hA:1

Short vowels (followed by a long consonant):

Symbol	Word	TRANSCRIPTION
I	vitt	vIt
е	vett	vet
Е	rätt	rEt
Y	bytt	bYt
u0	\mathbf{buss}	bu0s
2	föll	f21
U	bott	bUt
0	håll	hOl
a	hall	hal

There are also two pre-r-allophones (long and short) of /E/ and /2/ (see below).

The following important allophonic variants occur in Swedish which require separate symbolic representation:

\mathbf{Symbol}	Word	Transcription	
{:	här	h{:r	pre-r allophone of E:
9:	$\mathrm{f\ddot{o}r}$	f9:r	pre-r allophone of 2:
{	herr	$h{r}$	pre-r allophone of E
9	förr	f9r	pre-r allophone of 2
0	pojken	pOjk@n	schwa vowel allophone
rt	$_{ m hjort}$	jUrt	retroflex consonant, not ini-
			tial
rd	\mathbf{bord}	bu:rd	"
rn	\mathbf{barn}	bA:rn	"
rs	fors	fOrs	22
rl	karl	kA:rl	22

B.4 Levels of annotation and extension of SAMPA

B.4.1 SAMPA as a phonemic system

The present SAMPA system, which was provisionally agreed at the end of the Extension Phase, is defined as a system for phonemic transcription and annotation. This means that the symbols are used according to the analysis of distinctive sound oppositions within each language. Thus, although their relation to sound category symbols of the International Phonetic Alphabet (IPA) is given, they are symbols of intra-language convention, and do not have an exact language-independent phonetic (auditory or acoustic) equivalence, nor do they represent a single sound within a language.

For example, the symbol /t/, used in the transcription of all 8 partner languages, could represent an unaspirated sound in French or Italian, a strongly aspirated sound in German or English, and an affricated sound in Danish. In English the /t/ can also stand for an unaspirated sound (following /s/) or the more usual aspirated sound. Vowel symbols often represent widely diverging sounds from one language to another; /{/ in Danish is very different from /{/ in English, for example.

This basically phonemic, or sound-system-orientated (systematic) function of SAMPA means that a general extension of the SAMPA coding system to allow fine phonetic differentiation of speech sounds is not possible. There are, however, examples in the SAMPA list of symbols which can be used to represent non-distinctive differences within a language, e.g. "r" and "R" for regionally dependent free variants, and some important allophonic variants are allowed for (e.g. in Swedish and Norwegian). Also, auditory transcription (French "notation") is meant to be a "broad phonetic" representation of the actual utterance, including elisions and assimilations (inasfar as these can be represented with the phonemically orientated SAMPA inventory) rather than the strictly phonemic string of the citation form.

One area in which an extension of SAMPA is possible, indeed probable, is prosody. Certain "Boundary and Prosodic Features" have been agreed preliminarily, but their use has only been illustrated in the English EUROM-0 transcriptions. The considerations of prosodic description in a multilingual context may well reveal the need to modify and extend SAMPA. The work on prosodic description may also conclude that a separate prosodic annotation tier is necessary.

B.4.2 Detailed phonetic or acoustic annotation

For finer segmental annotation of speech recordings, three basically different approaches are offered for discussion. All three approaches require a separate annotation tier, but the labels are temporally defined by the location of the phonemic segment boundaries (phonemic markers in the case of centre labelling).
1. The SAMPA symbols are given language-independent sound values (IPA equivalent values) and modified by means of agreed diacritic codes to reflect fine phonetic detail.

Advantages:

• No new segmentation or marker placements would be required.

Disadvantages:

- Different symbols would sometimes be required at the phonemic and phonetic levels, particularly for vowels. For instance, Danish /{/ might have to be represented by phonetic [E]; English /{/ might have to be represented by phonetic [a] or even [A], depending on regional accent.
- Diacritic symbols would have to be agreed for all partner languages.
- ASCII coding on one keyboard would possibly not be sufficient for the necessary IPA symbols and diacritics, and there would be little or no mnemonic value in the choice of many symbols.

2. The SAMPA phonemic values are retained for each language, and the phonemic segment is subdivided into acoustically quasi-homogeneous elements. For example, /k/ may contain a partially voiced closure, a clear burst, and a period of aspiration prior to the vowel onset. Note that this approach is an acoustic-event labelling and is used in a similar way at CERFIA, IES and UCL. The following characterisation retains the primary symbol as "pointer" to the phonemic identity of the utterance:¹

- kv = Voiced portion of closure
- kc = Voiceless portion of closure
- kp = k-burst
- ka = k-aspiration

Advantages:

• The acoustic realisation of each phonemic segment is defined in greater detail than is possible even in narrow phonetic transcription, where, for example, a partially voiced closure cannot be easily represented.²

Disadvantages:

• New segment markers have to be set.

 $^{^1}$ The symbol combinations used here are only examples, taken from a system used at UCL. They merely serve to illustrate the type of fine acoustic categories under discussion. Other symbols or symbol combinations could be and have been used by others to represent similar categories.

 $^{^2}$ Suggestions for the transcription of pathological speech made at the Kiel IPA Convention envisage this sort of detail, but they are still under discussion and have not yet been considered with respect to computer transcription.

• The system can only apply to approaches that recognise the need to define segment boundaries (however arbitrary they may be theoretically).

Note: It must be pointed out that the two-symbol representation given above is redundant, in that the acoustic-event categories are common to phoneme classes rather than individual phonemes; i.e. pc, tc, and kc would all be a period of voiceless closure and therefore not require the place specification. Also, if the phonemic category is specified in a different tier of annotation, it is recoverable, and may be used for a database search, e.g. with a view to developing a set of rules covering the possible "internal" structures of stretches of signal associated with a particular phoneme. At present, some partners need to retain the "phonemic pointer" in order to derive the phonemic label file from the lower level acoustic-event file.

3. A third approach, favoured by the linguistic group at ICP (Grenoble) recognises transitional phases between areas marked as optimally representative of a particular phoneme category. The finer labelling requires the delimitation of the (centre-marked) optimal area, thus also delimiting the area of coarticulation.

Advantages:

• The theoretically doubtful "changeover point" from one "phoneme" segment to another is avoided, and areas of indeterminacy are identified.

Disadvantages:

• New markers have to be set.

Each of these approaches would provide an annotation which is closer to the (acoustic-) phonetic realisation of the utterance than the phonemic SAMPA labels. For the development of speech knowledge in general, and for the definition of rules describing the structure of continuous speech in particular, the use of a more detailed annotation is essential. It is the symbolic bridge between measurable acoustic parameters and abstract phonological categories. Which approach is selected for more detailed annotation within the SAM project depends on the use to which it will be put. Essentially, the closer a symbolic representation comes to significant acoustic events (whereby "significant" is an application-dependent term), the more useful it will be in speech-knowledge acquisition and rule development. Both synthesis and recognition assessment can only gain.

	SAMPROSA	ASCII	Definition	
Local tone	Н	72	High pitch	
	L	76	Low pitch	
	Т	84	Top pitch (extreme H)	
	В	66	Bottom pitch (extreme L)	
	М	77	Mid pitch	
	+	43	Higher pitch	
	++	43,43	Much higher pitch	
	+-	43,45	Peak (upward-downward)	
	-	45	Lower pitch	
		45, 45	Much lower pitch	
	-+	45,43	Trough (downward-upward)	
	^	94	Upstep	
	~~	94, 94	Wide upstep	
	!	33	Downstep	
	!!	33,33	Wide downstep	
	= or > or S	$61 \ 62 \ or \ 83$	Level or same tone	
Terminal tone: from Local and Nuclear tone repertoires				
Nuclear tone	-	45	Level tone (before tone group boundary)	
	' or / or R	$39 \ 47 \ or \ 82$	Rising tone	
	' or \setminus or F	$96 92 { m or} 70$	Falling tone	
	'' (etc.)	$96,39~({ m etc.})$	Fall-rise	
	'' (etc.)	$39,96 ({\rm etc.})$	Rise-fall	
Length	:	58	Segment length mark	
Stress	"	34	Primary stress	
	%	37	Secondary stress	
Pause		46, 46, 46	Silence	
Boundary	\$	36	Syllable boundary	
	#	35	Word boundary	
		124	Tone group boundary (non-directional)	
	[91	Tone group boundary (left)	
]	93	Tone group boundary (right)	
Metasymbols	-	45	Separator (the underscore,	
			_, ASCII 95, may replace this owing	
			to ambiguity with level tone)	
	*	42	Conjunctor	

B.5 SAMPROSA (SAM PROSodic Alphabet): symbol set

For the full description see Document No. SAM-UCL-037: Wells, J., Barry, W., Grice, M., Fourcin, A., Gibbon, D., 1992, *Standard Computer-Compatible Transcription*. ESPRIT project 2589 (SAM). SAMPROSA was proposed by Dafydd Gibbon in consultation with colleagues in the SAM project and elsewhere.

SAMPROSA is designed as a superset of a number of symbol sets used to denote phonological and phonetic prosodic categories, for application in multi-tier transcription and representation systems. In such transcriptions and representations, independent parallel symbolic representations are made of an utterance (a signal) using different segmental or prosodic criteria. The parallel symbolic representations may be related in two different ways:

- 1. Association: Phonological rules are defined which assign prosodic units to segmental units as tags at a linguistic level of definition.
- 2. Synchronisation: The symbols may be assigned to the signal time function as labels or annotations; the temporal relations between symbols are then given empirically (extensionally) via their position with respect to the signal.

The main practical applications for SAMPROSA are in the areas of

- prosodic transcription for linguistic purposes, and
- prosodic labelling in speech technology and experimental phonetic research

B.6 References

SAM 1988. ESPRIT Project 1541: Definition Phase Final Report "Multilingual Speech Input/Output Assessment Methodology and Standardisation". London: University College London.

SAM 1989. ESPRIT Project 1541: Extension Phase Final Report. London: University College London. VI.2: First appraisal of SAMPA.

Wells, J.C., 1987. "Computer-coded phonetic transcription", Journal of the International Phonetic Association 17:2, pp. 94–114.

C SAM file formats

C.1 Introduction

To facilitate multi-laboratory exploitation of speech database material, SAM has defined a standard, transparent specification of all files used in the acquisition of post-processing of speech recordings for assessment and research purposes. In addition, a standard SAM recording software package, EU-ROPEC, developed at ICP – Grenoble, uses a set of SAM standard speaker and prompt file structures and produces speech and orthographic label files of standard format. This section gives details of formats and functions of files, and defines their function in the different stages of acquisition and processing.

C.2 Speech file and associated description file formats

It is now agreed as a standard for SAM speech databases, that a *speech file* contains only speech waveforms, and that an *associated description file* is generated at the recording session. Thus the files are matched, their names being identical, except for the last letter of the extension.

For example, if the speaker AA records the corpus number BB (list of six sentences in English), and the current available file number in the recording lab is nnnn, the files produced will be:

AABBnnnn.SES	sampled speech
(AABBnnnn.SEL	L for Laryngograph)
(AABBnnnn.SE2	for the second channel signal file)
AABBnnnn.SEO	associated description file generated automatically
	during recording.
	(O = orthographic time-aligned labelling)

The associated description file has standard *label file format*, with a header and a body. (see C.3.1 Header format for label files; C.3.2 for body of label file). It contains all the information usually needed by people working on the files without a database management system.

C.3 Label file format

A label file consists of a *header* and one or more *label bodies*. The header consists of the header keyword "LDH:" and a number of lines providing information about the labelling and the speech file to which it applies. The number of lines in the header is not fixed. The header continues until the label body "LBD:" keyword appears.

The label body specifies the type and location within the speech signal file of each segment that has been labelled. It continues until either another label body keyword "LBD:" or an end-of-label-file "ELF:"

BASIC LABEL FILE STRUCTURE:

The label header LHD: ... from LBD: to the label body from LBD: to ELF: or a new label body from LBD: ELF: to

C.3.1 Label file header format

LHD: header keyword + version (version V4.0, March 1991) FIL: file type TYP: specific file type (eg prompt orthographic, spoken orthographic, phonemic, prosodic etc) DBN: database name VOL: database volume ID DIR: directory (for the source file) SRC: source file name CMT: comment TXF: name of the textfile (comment on what this is supposed to contain - namely what was intended to be said, ie prompt file.txt or nothing (if it is a prompt or free speech label file then could be left blank) but could be a file containing instructions) CMT: comment SAM: sampling rate BEG: labelled sequence start position END: labelled sequence end position RED: recording date RET: recording time REP: recording place SNB: number of (8-bit) bytes per sample SBF: sample byte order SSB: number of significant bits per sample RCC: recording conditions code (define a set of values, micro type, position..) NCH: number of channels

```
SPI: speaker information: sex, age, native language
PCF: protocol file name (recording protocol used)
PCN: protocol number
CMT: comment
EXP: labelling expert
SYS: labelling system
DAT: date of completion of labelling
SPA: SAMPA version - note 2...and 3.... refer to different levels
CMT: comment
```

Note that the SAM labelling tool PTS requires all fields to have an entry otherwise it will not open the label file; "—" should be entered when the information is not available.

C.3.2 Label file body format

Between the label body keyword (LBD:) and the end-of-label-file keyword (ELF:) four categories of mnemonics can occur:

1. Label type:	LBR, LBO, LBB, LBP etc. (see below C.3.3)
2. EXT:	Line extension (In practice this is only used in or-
	thographic recording
3. CMT:	followed by any comment needed
4. DSC:	Indication of discontinuous recording mode, i.e. that
	pauses between utterances have been automatically
	excised during recording (see $C.3.4$ a) below)

C.3.3 Label types

- LBR: This label type occurs a special kind of label file which is created automatically during recording with EUROPEC (see 3 below). It contains sequence beginning (in samples), sequence end, input gain or recording, minimum sample value, maximum sample value, orthographic text. The text represents the prompt, not necessarily what the speaker actually uttered. In the case of two-channel recordings, each LBR label can be followed by an LB2: sequence (second channel speech file) or LBL: (laryngograph on 2nd channel) that contains the same information: beginning value (in samples), end value, input gain or recording, minimum sample value, maximum sample value. A mnemonic indicating other sensors (e.g. LBN (nasal), LBF (airflow), LBT (tongue contact) etc. can be defined if required. The purpose of this label file is to call up individual sections of larger files. The LBR labels represent "items" which can be specified when a signal file is opened.
- LBO: Orthographic labels produced manually. These labels attempt to represent what the speaker actually produced, with indications of pauses, hesitations, repetitions etc. Often, an EXT: line and a CMT: line are needed.
- LBB: Broad Phonetic labels, produced manually or by means of (semi-) automatic label alignment.

- LBA: Acoustic-phonetic labels (subdivisions of broad phonetic segments) produced manually or by means of semi-automatic labelling.
- LBP: Prosodic labels, produced manually or by means of (semi-) automatic labelling.

C.3.4 Label File example

a) This is file DFS20014.SEO, created automatically during a single-channel recording of file DFS20014.SES.

```
LHD: V4.0
FIL: label
TYP: prompt orthographic (comment needed that mixtures are possible)
DBN: -
VOL: EUROM.1
DIR: ENGLISH
SRC: DFS20014.SES
TXF: S2.TXT
CMT: Information about the recording session
SAM: 16000
BEG: 0
END: 431872
RED: 07/11/89
RET: 15:10:33
REP: I.C.P. Grenoble (FR)
SNB: 2
SBF: 01
SSB: 16
RCC: 1
NCH: 1
SPI: M, 39, French
PCF: SENTEN.DES
PCN: 1
CMT: Information about the labelling session
EXP: -
SYS: -
DAT: -
SPA: -
CMT: Item: label start, end, input gain, min level, max level, string
LBD: -
LBR: 0, 55551, 0, -5128, 4775, Decimal numbers are an aid in
EXT: adding up.
DSC: -
LBR: 55552, 158975, 0, -7680, 8878, Monetary systems have
EXT: evolved to make use of this base ten notation.
DSC: -
LBR: 158976, 223743, 0, -7123, 7562, France became the first
EXT: decimal country in Europe.
```

```
DSC: -

LBR: 223744, 275199, 6, -12487, 13262, Germany's decision

EXT: followed eight years later.

DSC: -

LBR: 275200, 361983, 6, -11965, 12451, Scandinavian States and

EXT: Russia changed in eighteen seventy-five.

DSC: -

LBR: 361984, 431872, 6, -12902, 14320, Britain chose to have

EXT: decimal money only in nineteen seventy-one !

ELF: -
```

```
b) This is file ADS10010.SFO, the orthographic label file created automat-
ically during the two channel recording of ADS10010.SFS
LHD: V4.0
FIL: label
TYP: orthographic
DBN: EUROM_1
VOL: -
DIR: -
SRC: ADS10010.SFS
TXF: S1.TXT
CMT: Information about the recording session
SAM: 20000
BEG: 0
END: 193023
RED: 11/Apr/90
RET: 16:47:46
REP: ICP
SNB: 2
SBF: 01
SSB: 16
RCC: 2
NCH: 2
SPI: M, 49, French
PCF: PEQPHRAS.DES
CMT: Information about the labelling session
EXP: -
SYS: -
DAT: -
SPA: -
CMT: Item: label start, end, input gain, min level, max level, string
LBD: -
LBR: 0, 17663, 0, -12382, 14455, Maman a préparé une galette
EXT: pour jeudi ?
```

```
LB2: 0, 17663, 0, -12384, 14439
DSC: -
LBR: 17664, 41215, 0, -6954, 9023, Ces élèves prendront
EXT: l'autocar tout à l'heure !
LB2: 17664, 41215, 0, -6958, 9024
DSC: -
LBR: 41216, 86527, 0, -11750, 15336, Parfois, mon épicière
EXT: vend à crédit.
LB2: 41216, 86527, 0, -11756, 15348
DSC: -
LBR: 86528, 137983, 0, -11809, 11572, Personne n'a applaud i ce
EXT: beau discours ?
LB2: 86528, 137983, 0, -11816, 11560
DSC: -
LBR: 137984, 193023, 0, -15097, 18881, Je me demande pourquoi
EXT: on court sans cesse.
LB2: 137984, 193023, 0, -15102, 18880
ELF: -
```

C.4 Files used in a recording session using EUROPEC

The standard SAM recording software EUROPEC, is geared to producing signal and label files of standard format. In turn the software requires input data which is formatted according to the following specifications.

C.4.1 Corpus File: CORPUS.DBF

The Corpus File lists all the prompt files used in a particular set of recordings. This file is scanned whenever a file is specified for a recording "take". The prompt files, which have a 2-character name and extension .txt, are arranged in an alphabetical catalogue A–Z, each letter containing up to 36 prompt files (e.g. A0–A9 plus AA–AZ). Data entries for the prompt files are of two types:

1. those required for the prompting process, and

2. those required for the file header.

Order of data entries:

```
CMT: Optional entry providing information about the prompt file,
    typically the prompt file name and a short description.
CCD: Corpus code ( 2 letters (ex: P2) )
DBN: Database name (20 characters max)
CNM: Corpus name (40 characters max)
CTY: Item type (one letter)
NBI: Number of items in the file (int)
```

LAN: Language (char) PCF: Protocol Description File (XXXXXXX.DES) _____ _ _ A ------. _____ Ρ CMT: P1.TXT Prompt text for passage 1 CCD: P1 DBN: EUROM_1 CNM: A reading passage (ESPRIT CD) CTY: P NBI: 1 LAN: E PCF: PASSAGE.DES -----Q _ -----S -CMT: S2.TXT Prompt text for Block 2 of 5 sentences CCD: S2 DBN: EUROM_1 CNM: Sentences (UCL) CTY: S NBI: 5 LAN: E PCF: SENTEN.DES -----Т _ CMT: T3.TXT Prompt text for Block 3 of 30 digit triples CCD: T3 DBN: EUROM_1 CNM: list 3D Digits Triples CTY: N NBI: 30 LAN: E PCF: DIGIT.DES -----

738 Part IV: Spoken language reference materials

C.4.2 Speaker File: SPEAKERS.DBF

This file contains details about the speaker which can be accessed via the speaker codes. To maximise the speaker number, arbitrary code allocation AA–ZZ (=680) should be used rather than the speaker initials used in the examples. In many countries, the SNM and SBN lines will be left blank for data protection purposes.

Description of speaker specificities

```
_____
order :
SCD: Speaker code (2 char)
SNM: Speaker name (75 char max)
SBN: Speaker birthname (75 char max)
SEX: Sex (one letter)
DOB: Date of birth (year)
HET: Height (metres)
WET: Weight (kg)
NLN: Native language (75 char max)
ACC: Accent "
ETH: Ethnic group "
EDL: Education level "
SMK: Smoking habit "
PTH: Pathology "
_____
A
_____
D
SCD: DJ
SNM: DURAND
SBN: Joseph
SEX: M
DOB: 1946
HET: 1,73
WET: 70
NLN: French
ACC: Ardèche
ETH: white
EDL: -
SMK: heavy smoker
PTH: -
-----
SCD: DS
```

SNM:	DUPUY
SBN:	Simone
SEX:	F
DOB:	1952
HET:	1,60
WET:	58
NLN:	French
ACC:	-
ETH:	white
EDL:	-
SMK:	-
PTH:	-
• • • •	••
Е	
-	
• • • •	
-	
Z	
-	

C.4.3 Prompt file

The prompt file is given the name of the Corpus Code (CCD:) specified in the Corpus File. The name corresponds to its position in the Corpus File (X2 signifies that it is specified in 2nd position under letter x).

- Each prompt text file begins and ends with a specific mnemonic "SPR: EPR:".
- Each item is preceded by the mnemonic "TXT:".
- Each line that is a continuation of the preceding one, begins with the mnemonic EXT.
- If the text file is to be used in continuous recording mode, it must specify the display time in seconds for each item (following the mnemonic "DLA:").

EXAMPLES OF PROMPT TEXT FILES

S1.TXT (5 sentences) S1 is the corpus code

SPR: (start prompt file)
TXT: I have a problem with my water softener.
TXT: The water-level is too high and the overflow keep dripping.

740 Part IV: Spoken language reference materials

```
TXT: Could you arrange to send an engineer on tuesday morning ?
TXT: It's the only day I can manage this week.
TXT: I'd be grateful if you could confirm the arrangement now.
EPR: (end prompt file)
S2.TXT (5 sentences) S2 is the corpus code
SPR:
TXT: Please put me through to the complaints department.
TXT: The repair to the water main outside my house was
EXT: unsuccessful, and my cellar's flooded.
TXT: Your Water Services Department was singularly unsympathetic.
TXT: All their repair teams are apparently booked out for the
EXT: next two weeks.
TXT: Am I supposed to use the cellar as a swimming pool till
EXT: then ?
EPR:
Note that if a sentence is more than one line long, there will be one
extension line (field EXT).
P1.TXT (passage)
                        P1 is the corpus code
SPR:
TXT: Please put me through to the complaints department. The
EXT: repair to the water main outside my house was unsuccessful,
EXT: and my cellar's flooded. Your Water Services Department was
EXT: singularly unsympathetic. All their repair teams are
EXT: apparently booked out for the next two weeks. Am I supposed
EXT: to use the cellar as a swimming pool till then ?
EPR: -
Note that all this passage will be displayed as one item (only one TXT
field and several EXT fields).
ST.TXT (5 sentences)
                            ST is the corpus code
To be recorded in "continuous mode" (= all pauses between sentences are
retained).
SPR: -
TXT: Please put me through to the complaints department.
DLA: 3.0
TXT: The repair to the water main outside my house was
EXT: unsuccessful, and my cellar's flooded.
DLA: 5.0
TXT: Your Water Services Department was singularly unsympathetic.
DLA: 3.5
```

```
TXT: All their repair teams are apparently booked out for the
EXT: next two weeks.
DLA: 4.0
TXT: Am I supposed to use the cellar as a swimming pool till
EXT: then ?
DLA: 3.8
EPR: -
```

C.5 Protocol description file

In a recording session, the prompt presentation is driven by a specific file called Protocol Description File (specified after the PCF: Corpus File entry). It specifies the various stages and messages of the prompt presentation.

C.5.1 Codes used for presentation protocol (Protocol Description File)

```
Syntax is S1: <Field_ID>:<value>
       or S2: <Field_ID>:<value>, <value>, <value>, ...
CLS: ClearScreen and put cursor to 0,0
VON: Video attribute ON: the following attributes are activated (S2)
Authorized attributes are: center : text will be centered
                            blink
                                     : blinking mode
                            inverse
                                     : video inverse mode
                            underline : text will be underlined
                            bright
                                      : text will be bright
VOF: Video attribute OFF (S2)
BEL: ring bell
MSG: message to be displayed according to valid video attributes
NWL: newline on screen
PAU: n (integer) = duration of the pause in seconds
TRN: set training mode, no recording
CMT: comment
RWD: rewind external file (corpus text file)
GET: n (integer)
     n: number of items to be read in the external file (corpus text
file) then displayed and recorded (if recording mode is activated) or
not (if training mode is activated)
DTA: (string) Data. The item to be displayed (and recorded or not) is
given here.
EXT: extension to previous line.
SKP: n (integer)
     n: number of lines to be skipped in external file (corpus text
        file)
LOP: -n, m loop in the present file
     n: number of one-line-back in this file
     m: number of iterations
```

C.5.2 Example of Protocol Description File

This is SENTEN.DES (Protocol Description File for a corpus like S1 or S2 and given after PCF: in the Corpus File entry)

```
SPF: start of protocol file
CLS: -
VON: inverse, center
BEL: -
NWL: 2
MSG: ENGLISH SENTENCES
NWL: 2
MSG: I.C.P. GRENOBLE
VOF: inverse
PAU: 2
CLS: -
NWL: 2
MSG: Let's practise a little :
NWL: 2
MSG: please read the following
NWL: 2
MSG: two sentences
PAU: 3
CLS: -
TRN: -
CMT: DISPLAY BEGINNING - TRAINING PHASE
RWD: -
GET:* 1
SKP:* 2
LOP: -2, 1
CMT: DISPLAY END - TRAINING PHASE
RWD: -
CLS: -
NWL: 2
```

MSG: Now you're going to be recorded NWL: 2 $\ensuremath{\texttt{MSG}}\xspace$: this session consists of five sentences NWL: 2 MSG: read each of them NWL: 2 MSG: while they are being displayed PAU: 4 CLS: -NWL: 8 MSG: CAUTION ! SESSION BEGINNING VON: blink NWL: 1 MSG: Hit Space or Click Left Button RET: -VOF: blink CMT: DISPLAY BEGINNING REC: -CLS: -GET:* 1 LOP:* -1, 4 CLS: -NWL: 2 CMT: END DISPLAY CLS: -NWL: 8 VOF: center VON: blink MSG: STOP VOF: blink MSG: Recording PAU: 1 VOF: inverse VON: center CLS: -BEL: -NWL: 4 MSG: It's all over for this file NWL: 1 MSG: THANKS NWL: 8 EPF: end of protocol

NOTE: For the ST corpus, using continuous mode, the protocol file is nearly the same, but:

• No LOP instruction is allowed.

```
744 Part IV: Spoken language reference materials
```

• There is one GET instruction for all the items to be recorded in sequence together.

C.6 Recording conditions file

A Recording Conditions File is selected (and can be modified) prior to each recording session. Some of the entries are accessed for completion of the label-file header, others are for general reference. A number of files specifying different recording conditions can be stored and selected when requested. It describes a referenced set of recording conditions. Example:

 $1.{\rm RCD}-1$ is recording conditions code 1, implying that this file is the first of a number of defined conditions.

SCD:	start of conditions	
RCC:	1	recording condition code
VER:	V3.0	label version
VOL:		needed for label file
DIR:		needed for label file
SNB:	2	sample byte number
SBF:	01	sample byte order
SSB:	16	sample significant bits
NCH:	1	number of channels
LGG:	0	laryngograph used or not
PCN:	1	protocol code number
SAM:	20000	<pre>sampling frequency (20kHz is {\sc Sam}</pre>
		standard for EUROM.1ff)
MIN:		micro name
MIT:		micro type
MIP:		micro position
MID:		micro distance
NOB:		noise bandwidth
NOL:		noise level
CHB:		channel bandwidth
CHN:		channel noise
FLT:		additional filter
PRO:		external processing
ENV:		environment
ECD:		end of condition

C.7 PRIVATE.PEC file

This file is local to the SAM workstation. It provides information used for the recording process (OROS board type and address) and for completion of the label file header, as well as registering the number of files recorded

on the workstation.

Example:

BTY: 2Oros board type (1: AU21 2: AU22)ADR: 784Oros board addressLOC: I.C.P. Grenoble (FR)Recording placeNUM: 0000First file numberNSC: 1Number of monitors

SAM 1992. "File formats", in: ESPRIT Project 2589, Multi-lingual speech input/output assessment, methodology and standardisation, "User guide to ETR tools", Ref: SAM-UCL-G007, University College London, pp. 15–28.

D SAM recording protocols

D.1 Definition of terms

- SAMPLE A digital representation of the analogue waveform at any one instant in time
- SAMPLING RATE The (normally constant) rate at which SAMPLEs are collected and transferred to a storage medium.
- SAMPLING INTERVAL The time between successive SAMPLES fixed for any one SAMPLING RATE while the sampling process is not stopped.
- DISCONTINUITY The point in a stream or stored set of SAMPLES where the sampling process has been stopped for longer than one SAMPLING INTERVAL.

(Explanation – it is equivalent to the splice edit point on the tape of an analogue tape recorder.)

SEGMENT A set of sequential SAMPLES held on the storage medium which does not contain a DISCONTINUITY. A SEGMENT will be bounded by two DISCONTINUITIES

(Explanation – this means that during the period over which a SEG-MENT has been collected the process of sampling and storing has not been stopped.)

SPEECH FILE A set of sequential SAMPLEs held on the storage medium. The SPEECH FILE will contain one or more SEGMENTS.

(Explanation – the sampling process may have been stopped, then started again during the period over which a SPEECH FILE has been collected.)

TAKE The time over which a subject is invited to complete the production of a specific set of utterances. The TAKE will be recorded in one or more SEGMENTS. A SPEECH FILE will often (but not always) contain one TAKE.

(Explanation – the start of a single TAKE will often be identified by the issuing of a set of instructions to the subject and/or the inputting into a prompting system of the utterance set ID. It will normally be terminated by the prompting system finishing an utterance list and further operator-subject dialogue. There will normally be no operator-subject dialogue within a TAKE.

SESSION The period of time from when a specific subject arrives at the recording location to when he leaves. During this time one or more TAKEs will be recorded.

SPEECH CORPUS A set of SPEECH FILES.

SPEECH DATABASE One or more SPEECH CORPORA complete with the information to allow the CORPORA to be usefully accessed. This information will usually be in the form of time-aligned annotation, speaker descriptions, recording condition files, etc.

- 748 Part IV: Spoken language reference materials
- D.2 Classification of general strategies for recording and prompting

When designing a SPEECH DATABASE a number of decisions have to be made concerning the size and content of utterance lists, size of TAKES, length of SESSIONS, methods of prompting and recording, and dealing with errors. It is assumed in the following that a digital storage method is to be used, although many of the points in the following would apply in a slightly modified form for an analogue recording system.

There follows a classification and discussion of some of the design areas.

D.2.1 Recording mode

The RECORDING MODE defines the exact method of controlling transfer of speech samples to the storage medium. It is defined in terms of the TAKE, SEGMENTS and SPEECH FILES. It is concerned with the stopping and starting of the sampling process and the way the recording system treats any possible speaking errors.

MODE 1 A TAKE is recorded in one complete SEGMENT, errors are not excised.

(Explanation – this means that the sampling and transfer process is started at the beginning of the TAKE, all the acoustic signal, including breath noise, speaking errors are recorded and the sampling and transfer process is only stopped at the end of the TAKE.)

- MODE 2 A TAKE is recorded in one complete SEGMENT, except when a speaking error is detected during the TAKE, and then that error is excised from the storage medium. If there have been no speaking errors there will be no DISCONTINUITIES in the SPEECH FILE. For each section of speech removed due to a speaking error there will be one DISCONTINUITY.
- MODE 3 A TAKE is recorded in a number of SEGMENTS. There will be one SEGMENT per utterance, and speaking errors will be included in the SPEECH FILE.

(Explanation – in general this means that an endpoint detector is used to start and stop the sampling and transfer process. Apart from the "back-off" time no inter-utterance acoustic signal will be stored in the SPEECH FILE.)

MODE 4 A TAKE is recorded in a number of SEGMENTS. Speaking errors will be excised from the storage medium, and so there will be one SEGMENT per required utterance. This MODE is the same as MODE 3 except that the resulting file will contain no errors.

D.3 Prompting style

The PROMPTING STYLE primarily defines how a prompting system reacts when speaking errors have been detected during a TAKE and how the subject is instructed to behave after the production of speaking errors. The following classifications assume an automatic prompting system, but nonautomatic systems can be classified in the same way. For each style there are three different timing strategies.

- STYLE 1 ABORT TAKE and re-record. The subject is instructed that if he realises that they have made a speaking error then they should indicate the fact to the operator. When the operator is aware of a speaking error the prompting and recording system are stopped by a suitable "escape" mechanism. This situation is indicated to the subject and the prompting system is started to re-record that TAKE.
- STYLE 2 BACKUP-ON-FLY n utterances. If and when a speaking error is detected during a TAKE, the prompting system is backed up to a point before the error and continues from there. What happens to the recording will depend on the MODE being employed.
- STYLE 3 TACK-ON-END correction. If and when a speaking error is detected during a TAKE the prompting system carries on without stopping until the last required utterance has been produced and then the subject is prompted for one or more utterances to correct for the speaking errors. The recording action depends on the MODE employed.
- STYLE 4 NO PROMPTING REACTION. The prompting system carries on producing prompts in the pre-determined order. The subject may or may not be asked to react by correcting the error in some way.

D.3.1 Timing strategy

- REGULAR The timing of the prompt is independent of the rate at which the subject is actually speaking. In its simplest form each prompt is displayed at regular predefined intervals which will depend on the type of utterance that is to be produced. Alternatively the interval may vary from prompt to prompt depending on the expected (not the actual) time to speak the prompted text, i.e. dependent on the number of words/phonemes in a sentence.
- ENDPOINT The timing of the prompt is totally controlled by the production of utterances. The display of each new prompt is controlled by the detection of the endpoint of the last utterance.
- MIXED The timing of the prompt is controlled by a logical combination of a predetermined interval and the endpoint of an utterance. The display of each new prompt is triggered by whichever is later of the predetermined interval or the endpoint. This means that if a subject is slow the prompting system will slow down, but there is a maximum rate of prompting even for the fastest speaker.

D.4 Recording protocol

D.4.1 Microphone

A single B&K half inch pressure microphone number 4155 will be used connected to a 2230 level meter, used as a microphone amplifier. If a record-

ing site proposes to use another microphone or preamplifier the alternative choice of equipment should be ratified by other members of the project. IES has suggested a 4165 microphone, 2660 pre-amp and 2636 amplifier. Bochum University recommend the ADG C-414 microphone in omnidirectional mode connected to the John Hardy M-1 mic preamp.

The audio output of the level meter will be connected to both the line input of the OROS AU21 board and the left channel of the digital audio safety backup recorder. The method of splitting the signal will be made in such a way to ensure no degradation of the audio signal particularly in terms of loading, balanced/unbalanced connections and hum loops. Bochum University recommend the Brooke Siren Sytems microphone splitter.

The microphone will be positioned for 90 degree incidence, 50cm from lips 15 degrees off axis. (TNO configuration)

Tests by NPL have shown that this distance is acceptable in terms of speech signal to ambient/monitor noise.

The microphone position relative to the head/lips will be set and maintained by a headrest (or the not so consistent method of requesting the subject to maintain head position).

D.4.2 Other sensors

A Laryngograph will be used for some speaker/material combinations as specified in material choice section. The neck sensors will be positioned in accordance with UCL instructions, namely on either side of the point of the thyroid cartilage. The Lx output of the Laryngograph will be connected to the righthand channel of the digital audio (safety backup) recorder.

D.4.3 Speech data capture

Single channel, digital, direct recording to disc will be used for the microphone signal. Digital audio tape recorder safety backup will be made of microphone channel at all times. Second channel recording direct to disc and/or digital audio tape recording of the Laryngograph signal will be made as required.

The mechanism for direct digital recording to disc will be by use of the OROS AU21 board for single channel or an OROS AU22 for two-channel recordings on the SESAM workstation with EUROPEC software. The sampling rate will be 20 kHz, and the standard OROS digital filtering is used with an oversampling factor of 4. The nominal resolution of this system is 16 bits. Care should be taken to ensure that OROS board is placed in a slot in SESAM which minimises PC electrical interference.

The "standard" analogue filter on the OROS board is a 20 kHz low-pass filter with 0.3 dB ripple in the passband, 160 dB/octave slope and 80dB rejection. This is followed by a digital filter implemented on the TMS320C25 having the -3dB point at 8kHz. Any DC offset is automatically removed by the

OROS/EUROPEC software and the gain of the line input will be set within EUROPEC on a "per speaker" basis.

The safety backup (and second sensor recordings if an OROS AU22 is not used) will be made on a digital audio tape system such as the SONY Video PCM or DAT systems, having a specification with regard to bandwidth, signal-to-noise ratio and wow and flutter in excess of that of the OROS AU21 board. The recorder shall be left recording for the WHOLE of a session and not be stopped or rewound at all during the session. This removes the chance of over-recording wanted portions, reduces the operator effort will allow later analysis of a number of features not captured on the direct to disc recordings, such as number of retakes, motivation level of subject, coughs etc.

D.4.4 Recording environment

The recordings will take place in an anechoic room.

The subject will not be provided with an electrical sidetone path for their own voice.

The lighting and temperature levels will be such as to cause no stress to the subject and care must be taken to ensure that there is no distracting light reflection off the prompting screen. Care must also be taken to ensure that the anechoic nature of the recording is not compromised by allowing significant acoustic reflections to be picked up on the microphone. The prompting monitor should be slightly angled and there should be no tables or other sound reflecting surfaces between the subject and the microphone (or tables to be covered with sound absorbing material).

D.4.5 Recording mode and prompting style

It has been agreed that the IASM-A CORPUS should contain no speaking errors. Further the inter-utterance pauses should be captured in their entirety and there should be no DISCONTINUITIES within a TAKE.

PROMPTING STYLE 1 WILL BE USED TOGETHER WITH RECORDING MODE 1. THERE SHALL BE ONLY ONE TAKE PER FILE.

The recording system will be integrated with the prompting system and this is to be as automatic as possible (see EUROPEC Section III below). It is clearly possible for the prompting system to start the recording system before the first prompt. In the general case of RECORDING MODE 1 as discussed in the earlier part of this document, the end of the final utterance is not defined. However, it has been agreed that a base level annotation will be automatically produced during the TAKE. This will consist of the prompted orthographic text along with the endpoints of each utterance.

As each endpoint is known to the prompting/recording system, then this information can be used to stop the recording process at the end of the final utterance.

A TIME ALIGNED ORTHOGRAPHIC ANNOTATION WILL BE PRODUCED AUTOMATICALLY DURING THE TAKE -THE RECORDING PROCESS WILL BE STOPPED AUTO-MATICALLY AFTER THE FINAL UTTERANCE.

However, if in the light of experience, it is found that significant SESSION time is wasted in re-recording TAKES or subjects are being excessively stressed by the "ABORT TAKE on error" style then the next most appropriate technique would be to "BACKUP-ON-FLY" and excise any errors from the TAKE. This technique will require an addition to the previously agreed standard of annotation file, namely a label to indicate the position of the DISCONTINUITY so caused.

PROMPTING STYLE 2 AND RECORDING MODE 2 WILL BE USED ONLY EXCEPTIONALLY AND WITH THE AGREEMENT OF OTHER MEMBERS OF THE CONSOR-TIUM – A DISCONTINUITY MARKER TO BE ADDED TO THE ANNOTATION FILE. ONE TAKE PER FILE.

It has been suggested that a regular timing strategy puts too much stress on some speakers and that there is evidence of excessive speaking rate when endpoint timing (i.e. speaker driven prompting rate) is used.

MIXED TIMING STRATEGY WILL BE USED

The exact value for tP (the minimum prompting interval in this strategy) depends on the display time (DLA: [time in seconds] in the prompt file) plus the stop back-off time.

D.4.6 Recording control

The subject will be prompted by an automatic system – EUROPEC running on the SESAM workstation positioned outside the anechoic room. The EUROPEC system will be controlled by a recording manager (operator), viewing the standard SESAM monitor. The subject will use a second monitor placed in the anechoic chamber, taking its video input from a "T" connection from the SESAM video output. The size and positioning of the monitor will allow the subject to see the prompted text and level meter without stress. The positioning of the monitor with respect to the microphone will ensure that there is no detectable noise pickup, either acoustically or electrically (particularly at line and frame frequencies.) Bochum University suggest that by using EGA or VGA graphics mode the line frequency is above 20 kHz and therefore unlikely to be a source of interference.

The speaker to microphone distance, azimuth and elevation will be controlled by ensuring minimal head movement commensurate with minimum subject stress. A headrest is highly recommended.

The subject will have, for the material/country combinations specified in the recording material section, an acoustical prompt generated by a second SESAM fed by the RS232 output of EUROPEC.

The operator will continuously listen to the subject at all times via the recording chain. This will allow subject-operator communication and enable the operator to do a 100% check on the speech material content. Bochum University suggest that 2 operators should be employed simultaneously to reduce fatigue. The headphone output of the PCM/DAT digital audio safety backup machine would be a suitable source for audio monitoring.

The operator will speak to the subject in the anechoic room via a one way intercom/loudspeaker system, which will be switched off during takes. The gain of the intercom system shall be set to a level at the ear representing speech 1 metre from the lips of an average talker. The setting of this level is important as it will tend to effect the speaking level of the subject – and should be consistent across recording sites.

Speaking effort will be controlled for some speaker/material combinations as specified in the reference material section. This will be done by the use of a speech level meter displayed on prompting monitor by EUROPEC. There will be three modes of operation:

- 1. The level meter gain set to nominal, after the speaker has stabilised in level the recording gain will be adjusted so that the normal peak level of the speech reaches a reference point, 12 dB below peak. The speaker is controlled for consistent speaking level during the subsequent take by being asked to keep to reference point as much as possible.
- 2. The level meter gain is set 6dB lower, the speaker keeps to reference point and so will speak 6dB louder.
- 3. The level meter gain is set 6dB higher than nominal, the speaker keeps to the reference point and so will speak 6dB more quietly.

Speaking style will be controlled by:

- 1. For some specified material the prompting system will show sets of sentences as a single item, and for other material they will be shown as separate items.
- 2. For some country/material combinations the speaker will be additionally prompted by an audio stimulus. The audio will be generated as a result of an RS232 serial output from EUROPEC/SESAM.

Speaking rate will not be controlled except that the maximum rate of presentation of prompted items will be at a rate set within EUROPEC dependent

on the material. This rate will be defined within the choice of material section.

There will be no attempt to produce stress in the subject. The total amount of material recorded in each session will be restricted to 8 minutes representing approximately a 30 minute elapsed time session. This limit will minimise stress on the subject and represents the maximum time that can be recorded on a 20 Megabyte disc.

The recording manager will ensure that the subject is in a relaxed condition. There will be no attempt to control any speaking effect due to the time of day of the session.

D.4.7 Recording procedure

At each session the speaker code of the subject will be confirmed. (Initially this will be done by checking a printed list of speaker details – an automatic system will subsequently be used.)

Session specific information will be entered via the EUROPEC system – some manually and some from the conditions file.

At each session the recording manager will provide an outline of the procedure for that session. This briefing should be in written or recorded form so as to achieve consistency between subjects.

At each session in the anechoic room the subject will be asked to practise speaking typical material for that session, with the level meter switched off. A note will be made of the peak level during the practice and the gain of the OROS AU21 will be set such that this peak level is 12–15 dB below maximum possible recording level. A large number of experienced recording sites have rejected the original suggestion of 6 dB head room. They feel that 12 to 15 dBs would be safer and reduce the possibility of takes having to be re-recorded due to overload, this in turn will reduce the stress on inexperienced speakers. The gain setting will be automatically recorded for inclusion in the session dependent recording conditions file.

If the session is to involve the use of control by level meter the subject should practise controlling their level with respect to the meter, for all (3) level meter gain settings.

The recording conditions/material etc. for each take will be entered into EUROPEC – this will cause the correct file names to be generated.

At each take the subject will, where appropriate, be given the opportunity to preview the prompt material. This is particularly needed for continuous speech material. In some cases, e.g. for particular forms of numbers, orthographically ambiguous nonsense material etc., additional instructions need to be given, and the subject given time for familiarisation with the task before the session.

The recording manager will monitor the speech production with reference to the prompted text. If there are any errors the take will be stopped and the subject asked to start the relevant take again. The disc file will be discarded and the same serial number reused for the re-recording. The digital audio tape recorder safety backup will not be stopped due to re-recording of errors – only at the end of a session. If after the integrity check a take has to be re-recorded then the digital audio safety backup recorder should be restarted.

D.4.8 Integrity checks

At the end of each session the quality of the recordings should be assessed by a 10% review, and the integrity of the item endpoint labelling should be 100% checked by using the VERIPEC module within the EUROPEC software.

D.4.9 Backup procedures

At the end of each session all files produced should be transferred to Exabyte 8mm tape system or via Ethernet to local mainframe for backup. The available disc space on the SESAM workstation is approximately 20Mbytes, representing 8 minutes of speech – this would therefore set a convenient and reasonable maximum to the recording session and represents approximately a 30–40 minutes elapsed time per session. All files produced during the session should also be backed up at the same time – the data is a valuable resource and files are easily lost. The digital audio safety backup tapes produced for each session shall be kept indefinitely as an archive of the session. It is expected that 4 to 6 sessions can be accommodated per tape, and care must be exercised not to over-record a previous session.

D.4.10 Retrieval procedures

If a speech file is lost AND its computer backup is also lost (OR the backup was never made!!!) then the digital tape recording safety backup will have to be used. This raises a number of questions. If the recordings are transferred to computer form in the digital domain, then an OROS AI-PCM type interface will be required – but this does not allow a resampling rate of $20 \, \text{kHz}$ – $19.8 \, \text{kHz}$ is closest possible. This retrieval method is only possible if the recording was made on a PCM or DAT machine where the pre-emphasis was switched off. The alternative is to copy in the analogue domain, the problems here are:

- 1. There is normally a degradation of frequency and phase responses and signal-to-noise ratio is reduced (the latter effect is possibly masked by the reported poor signal to noise of OROS A/D board).
- 2. The gain of the original signal is relatively poorly defined because the gain on the PCM unit is infinitely variable and there is no automatic method of noting the gain set for the microphone channel for each speaker. The calibration signal should be used for this purpose.

D.4.11 Calibration

The recording chain will be calibrated in two ways:

- 1. A B&K 1 kHz 4230 calibrator is placed over the microphone.
- 2. The injection of a 4:7 Mark-Space ratio square wave of 20 Hz and 100 Millivolts pk-pk. electrical signal into the microphone body. This is achieved by the use of:
 - 1. a signal generator (the circuit of which can be obtained from UCL);
 - 2. a B&K Input Adaptor, JJ 2614, which replaces the microphone capsule;
 - 3. a capacitor of a value equal to that of the capsule element replacing the 1nF capacitor in the B&K adaptor.

A recording of both calibrator signals shall be made every 20 sessions, providing confirmation of gain calibration, and recording path frequency response, phase response and noise figure. The length of recording of each calibrator shall be 10 seconds plus 10 seconds of no signal condition. It would be sensible to check the signal-to-noise ratio using these files at the site to confirm the correct working of the system.

The Laryngograph channel shall also be calibrated, using the UCL 4–7 mark-to-space ratio rectangular wave calibrator.

D.4.12 Inter site consistency and recording procedure verification

A pilot set of recordings will be made at each site consisting of 1 speaker speaking 10% of all the sentence and number material and all CVCs. This pilot material from each site will be transferred using floppy discs, one each for calibration, sentences, CVCs and numbers. The material will consist of the speech files together with corpus description files, labelling files, speaker file, text files and conditions file.

The floppy discs will be sent to NPL where the quality and consistency will be assessed by a panel of "experts". Modifications and clarifications to this set of protocols may then be made.

D.4.13 Collation of recordings

All the recordings at one site will be collated and checked. Checks will be made for:

- Quality of speech material and labelling on a 1% basis.
- Consistency of file names.
- Number of files of each type.

The central collation point will configure the material in the correct directory structure for the production of one or more sets of CD-ROM.

SAM 1992. "Recording protocols". In: ESPRIT Project 2589, Multi-lingual speech input/output assessment, methodology and standardisation, "User

guide to ETR tools". Ref: SAM-UCL-G007, University College London, pp. 5–14.

E SAM software tools

The following is a summary of the software packages which have been developed within the SAM project. With the exception of PTS (see Section E.4) and SAM_SPEX (see Section E.10), all software is freely available to other ESPRIT projects, but a charge will be made for media, postage and packing.

E.1 DKISALA (Interactive Semi-Automatic Labelling Software)

VERSION: 1.1, March 1992

HARDWARE / OS REQUIREMENTS: Standard SESAM workstation

- DESCRIPTION: The DKISALA software tool for semi-automatic labelling of speech corpora runs on the SESAM workstation. It is possible to listen to the speech signal during annotation if a non-SAM-standard PC speech processing DSP32C board is installed in the workstation (The AT&T development board or the Loughborough board PC system board based on the AT&T DSP32C signal processor. Details of configuration and board addresses can be given by the developing laboratory).
 - The DKISALA system uses a trained and calibrated Self-Organising Neural Network (SONN) to convert speech frame cepstrum coefficients into a set of continuously valued Acoustic-Phonetic Features, which are further transformed into a smaller set of Principal Components. The Principal Components are used to model individual allophones in a multivariate Gaussian probability density function, and these are processed by a Viterbi Search and Level-Building algorithm, which is constrained by the independently given string of phoneme symbols corresponding to the speech signal being labelled.
 - The SONN is trained on large speech corpora, which have been manually labelled prior to training. The presently used training data comprises EUROM-0 speech material from three speakers for the language under analysis.
 - The DKISALA system is presently working as a preliminary version of an interactive system. Results have shown that certain soundclass transitions are positioned very accurately and reliably, while some classes are regularly inaccurate.
 - The interactive component of the SALA system has been introduced to provide a means of preventing a cumulative error accruing as a result of a number of such sounds occurring within a short space. The automatic process stops when such a sound occurs. In such cases, the speech signal, the corresponding acoustic-phonetic features and the spectrogram are displayed on the graphic screen, and by using the mouse, the information on the screen and a listening facility, the user can propose the positioning of the specific boundary transition

to the system, which then takes control and reruns the label procedure. This procedure is repeated until the entire speech corpus is labelled.

DEVELOPING LAB:

Speech Technology Centre Institute of Electronic Systems Aalborg University Fredrik Bajers Vej 7 DK-9220 Aalborg, Denmark Tel: (45) 98 42 11 x 4866 Fax: (45) 98 15 67 40 CONTACT: Paul Dalsgaard

E.2 ELSA (ESPRIT Labelling System Assessment software)

VERSION: 2.4, December 1990

HARDWARE / OS REQUIREMENTS: SESAM workstation or UNIX workstation with X11 (e.g. SUN)

 $D{\rm Escription}:$

- The aim of this software is to provide the laboratories involved in the SAM project with a common and useful tool to assess semi-automatic labellers. It deals with the comparison between manual labelling of speech files (the reference) and semi-automatic labelling of the same files (both in the PTS format with SAMPA labels).
- Both files must contain the same phonetic strings. Therefore, the comparison can only be drawn in terms of boundaries location or centre location shifts. Included in this package are standard statistic tools, such as histograms, signed means and the best 10 and worst 10 boundary locations. It is possible to analyse either fine errors or gross errors and results can be written in a text file for analysis or printing. Confidence intervals are computed for a confidence level of 95%.

DEVELOPING LAB:

CRIN-CNRS & INRIA laboratories B.P. 239 54506 Vandoeuvre-les-nancy CEDEX France Tel: (+33) 83.91.20.00 Fax: (+33) 83.41.30.79 Email: {bourjot,boyer,fohr}@loria.crin.fr

CONTACT: Christine Bourjot, Anne Boyer, Dominique Fohr

E.3 EUROPEC (European Program d'Enregistrement de Corpus)

VERSION: 4.11, September 1991

HARDWARE / OS REQUIREMENTS: Standard Sesam workstation, with AU21/22 OROS board

DESCRIPTION:

- This is a corpus recording software which facilitates the recording of speech databases. It ensures constant recording conditions for various speakers and/or corpora. The recording session is driven by a protocol file (user-made) which describes the way in which items (constituents of the corpus) will be displayed on the screen.
- Features and functionalities:
 - "Standard SAM format" file production
 - One or two-channels
 - Recording and listening facilities
 - Three recording modes (manual, automatic, continuous)
 - Parameter adjustment (sampling rate, input and output gain, etc.)
 - Sound level meter
 - Free/operator controlled session
 - Error treatment
 - Easy data verification
 - File safeguards
 - Easy data retrieval and modification
 - Simplified recording session
 - User's guide, and many example files. User customisable in any language

DEVELOPING LAB:

I.C.P. (Institut de la Communication Parlee) 46 Avenue Félix Viallet 38031 GRENOBLE Cedex France Tel: (+33) 76 57 45 33 Fax: (+33) 76 57 47 10

CONTACT: Jerôme Zeiliger, Jean François SERIGNAT

E.4 PTS (Progiciel de Traitement de Signal)

VERSION:

- 4.40p (protected mode) March 1992
- + 4.40r (real mode) March 1992

HARDWARE / OS REQUIREMENTS: standard Sesam workstation with AU21/22 Oros board

PROTECTED MODE VERSION: at least 1 Mbyte of extended memory is required CONSOLE: MICROSOFT mouse, EGA or VGA display DESCRIPTION:

The tool consists of multi-window graphic software for speech:

- speech signal display; several formats including SAM standards;
- listening; total, selected records, between cursors, etc.;
- signal editing: copy, cut, move, etc.;
- speech processing: spectrogram, spectrum, etc.;
- manual labelling, semi-automatic label file correction;
- measurements;
- networking facilities (file transfers, remote login, etc.);
- user compliant by means of configuration files.

DEVELOPING LAB:

I.C.P. (Institut de la Communication Parlée)

46 Avenue Félix Viallet 38031 GRENOBLE Cedex France Tel: (+33) 76 57 45 39/33 Fax: (+33) 76 57 47 10

CONTACT: Jean-Claude Caërou, Jean-Marc Dolmazon

NB: This software package has different distribution regulations. The user should first sign a confidentiality agreement due to the use of two licenses:

- Rational System's DOS/16M extending memory for protected mode operations;
- Metawindow Plus graphic library.

E.5 RESAM

VERSION: 1.1, November 1991 HARDWARE / OS REQUIREMENTS: Low Level Sesam Workstation Description:

- The RESAM software package contains a program for changing sample rates of signal files, RESAMP, and a program for adding two signal files, ADD.
- The RESAMP program changes a signal's sample rate from a sample frequency Fa to a sample frequency Fb by up-sampling from Fa to the least common multiple of both Fa and Fb, and subsequently down-sampling to Fb. When up-sampling a signal k times, each sample is multiplied by adding (k-1) zeros. The up-sampled signal is interpolated using a low-pass filter and down-sampled m times by extracting each m-th sample.
- For example, when resampling noise signals (derived from Noise-Rom 0) from 20 kHz (up 4 times, down 5 times) into 16 kHz (Eurom

0 standard), a lowpass fir filter with a cutoff frequency of 16/2=8 kHz at a sample rate of $20^{*}4=80$ kHz has been defined (0.1 fs).

- Resampling a signal from 20 kHz into 16 kHz using the 1024 taps FIR filter will take about one hour per megabyte on a 286–12 mHz with co-processor. A program for designing 'windowed FIR filters' named WFIRDESN is also supplied. More details on the use of these programs can be found in the readme file on the distribution diskette.
- The program ADD allows to add signals with a specified gain factor and an offset value, which both can be determined with the SLM program. Also for the first signal, a time-offset measured in samples can be specified. This feature allows for a random start point in the noise files.
- Each i-th output-sample is calculated according to the following expression: Out[i] = Gain0 * (Input0[i + SampleOffset0] Offset0) + Gain1 * (Input1[i] Offset1) No indication is given if the 16 bits dynamic range is exceeded.

DEVELOPING LAB:

TNO Institute for Perception P.O. Box 23 3769 ZG Soesterberg Netherlands Tel: (31) 34 63 56 230 Fax: (31) 34 63 53 977 CONTACT: Jan A. Verhave

E.6 SAMITPRO (SAM Iterative Proportional Fitting)

VERSION: 1.1, March 1992

HARDWARE / OS REQUIREMENTS: Low Level Sesam Workstation Description:

- A software tool for analysis of interaction between recognition performance and test data characteristics. Examples of data characteristics are speaker characteristics such as sex, age, dialectal categories, educational level, weight, etc.
- Other types of characteristics could be recogniser type, testing site, microphone, testing environment, etc.
- All these examples are that they can easily be represented by subsets/sessions within the SAM_SCOR software, so that test performance can be automatically derived for each subset/session by means of SAM_SCOR.
- Subsets/sessions are to be defined in the SAM_SCOR control file which can easily be set up using the SAM tool RISE.
- Input to SAMITPRO is provided by applying the SAM_SCOR to a series of result files from recogniser experiment(s), so that the

performance is sampled in each cell of the input dimension.

- SAM_SCOR outputs the performance for each subset/session in a DBMS file, which can subsequently be read by SAMITPRO.
- SAMITPRO places the recognition performance in cells in a multidimensional table. Each dimension represents one characteristic, i.e. cells are described by a number of categories that are mutually exclusive. After filling up the table, hypotheses concerning the interactions of data characteristics (dimensions) can be evaluated by means of log-linear model fitting.

From this analysis conclusions can be drawn concerning the correlation between speaker characteristics and recognition performance: for instance whether some characteristics have an impact on performance or whether a complex interaction exists between several characteristics in combination. Further work aims at extending SAMITPRO to provide the capability of correlating recognition performance with speaker dependent speech inherent parameters derived by SAM_SPEX.

DEVELOPING LAB:

Speech Technology Centre Institute of Electronic Systems University of Aalborg DK - 9220 Aalborg, Denmark Tel: (45) 98 15 85 22 - 48 75 Fax: (45) 98 15 67 40 E-mail: oa@stc.auc.dk Fredrik Bajers Vej 7

CONTACT: Ove Andersen

E.7 SAM_REC0 Isolated Word Recogniser

VERSION: 1.0 November 1990 (corrections released January 1991)

HARDWARE / OS REQUIREMENTS: Standard SESAM workstation, equipped with OROS AU21 (or AU 22) DSP board and the library DISK_TRA of the recording software that accompanies OROS DSP boards

 $D{\rm ESCRIP\,TION}{\rm :}$

- The software implements an isolated word recogniser.
- A word is recorded and endpoint detection is then carried out by analysing an energy contour and the zero crossing rate.
- The endpointed word is preemphasised and windowed by a Hamming window.
- Discrete Fourier transforms are calculated over the word.
- The transform samples are filtered by a filter bank of triangular filters, linearly distributed in frequency up to 1 kHz and logarithmically distributed above 1 kHz.
- From the filter outputs cepstrum coefficients are calculated.
- The time warping uses eight cepstrum coefficients per frame, allows local time distortions between 50% and 200% and is based on a
symmetric slope constraint with P=1.

• The word chosen is the word among the reference templates with the smallest distance to the recorded word.

DEVELOPING LAB:

Telia Research Systems Research, Spoken Language Processing Rudsjoterrassen 2 S – 136 80 Haninge Sweden Tel: (+46) 8 707 55 72 Fax: (+46) 8 707 55 96 E-mail:gunnar.hult@haninge.trab.se

CONTACT: Gunnar Hult, Room A309

E.8 SAM_SCOR (Sam Input Assessment Scoring Software)

VERSION: 3.1, 3. March 1992 HARDWARE / OS REQUIREMENTS : Low Level Sesam workstation Description:

- SAM_SCOR is a software tool for scoring results from automated testing of recognition devices attached to the SESAM workstation.
- SAM_SCOR aims to cover the main classes of recognisers, from small vocabulary speaker-dependent isolated-word systems to medium/large vocabulary speaker independent continuous speech recognisers.
- For isolated word recognisers, standard statistics comprise correct recognition, miss, substitution, correct rejection and false alarm scores. Confusion matrices and similarity matrices are listed and the confidence for the conducted experiments is estimated. Special attention is given to the scoring of recogniser testing in which CVC type words are applied.
- Other features are:
 - a detailed listing of the different error conditions during testing (miss causes)
 - intergroup and intragroup scoring
 - the option of using the McNemar test for comparing recognisers that have been tested on the same database
- For connected/continuous speech recognisers, SAM_SCOR acts as a preprocessor for running the NIST string alignment and scoring program (Release 3.0, modified to run on SESAM, and distributed along with SAM_SCOR with the kind cooperation of David Pallett, NIST).
- Apart from the result files from testing a recogniser, SAM_SCOR needs as input a scoring control file in SAM file format, specifying the general conditions and type of scoring to be carried out. This

file may easily be produced manually (in a text editor) or be generated by applying the SAM software package RISE. The output of SAM_SCOR is both a standard documentation file, which is easy to interpret, and a DBMS file in SAM file format, which relates recogniser performance measures to speech material and speaker characteristics.

DEVELOPING LAB:

Speech Technology Centre Institute of Electronic Systems Fredrik Bajers Vej 7 DK – 9220 Aalborg, Denmark Tel: (45) 98 15 85 22 – 48 67 Fax: (45) 98 15 67 40 E-mail: bli@stc.auc.dk CONTACT: Borge Lindberg

E.9 SAM_SLM

VERSION: 2.0, 24 November 1991

HARDWARE / OS REQUIREMENTS: Low Level Sesam Workstation

- DESCRIPTION: The Speech Level Measuring program, SAM_SLM, has been developed in order to determine the speech level of a signal file. The method implemented for measurement of the signal level is based on a level distribution histogram of a squaring and lowpass filtering of the signal waveform, and the RMS value is calculated from this histogram. In addition the program permits the user to determine the DC offset of the signal file, too.
 - The program is self explaining, i.e. invoking the program without parameters results a display of usage. Even though SLM stands for Speech Level Measurement the program can be used to determine the levels of noise files as well.

DEVELOPING LABS:

 TNO Institute for Perception P.O. Box 23, 3769 ZG Soesterberg Netherlands Tel: (31) 34 63 56 230 Fax: (31) 34 63 53 977
 Jydsk Telefon Sletvej 30 DK - 8310 Tranbjerg Denmark Tel: (45) 86 29 33 66 Fax: (45) 86 29 90 68

CONTACT: Jeroen G. van Velden, Sven Danielsen

E.10 SAM_SPEX (Speech Parameter Extractor)

VERSION: 2.0, 25. Feb. 1992

HARDWARE / OS REQUIREMENTS: Low Level Sesam Workstation

Description:

SAM_SPEX provides an analytic basis for the definition of recogniser performance. The SAM project has adopted six speaker dependent speech parameter defined within the British Alvey project, MMI 132 STA. SAM_SPEX version 2.0 implements the extraction of these parameters from isolated words. The parameters measured are:

- Speaking Rate: The duration of each token is calculated using an energy-based endpoint algorithm (time resolution 1 msec) and related to the mean duration for all tokens of the same utterance.
- Energy: The ratio of the mean energy and the peak energy of the utterance.
- Pitch Frequency: The mean and variance of the pitch frequency estimated within the voiced parts of the endpointed utterance using a standard SIFT (Simple Inverse Filtering Tracking) algorithm.
- Voice Quality: The ratio of the energy between 0-2 kHz and 0-fs/2 Hz (fs = sampling frequency).
- Vocal Tract Area Distance: Calculated in two steps:
 - 1. The average area vector is computed.
 - 2. The vocal tract area distance is computed as the Euclidean distance between the average vocal tract area vector and the average of all vocal tract area vectors for tokens of the same utterance.
- Pattern Congruence: Calculated in two steps:
 - 1. The output values of an eighteen channels filter bank are computed every 10msec of a token.
 - 2. A dynamic time warping algorithm is used to calculate the distance between the token and all other tokens of the same utterance.

At present the definition of the parameters is under review and an improved definition is in preparation.

DEVELOPING LAB:

Jydsk Telefon Sletvej 30 DK - 8310 Tranbjerg - J Denmark Tel: (45) 86 29 33 66 Fax: (45) 86 29 90 68

CONTACT: Sven Danielsen

NB: This software is currently only available to SAM partners as the definition of the parameters is under review and an improved definition is in preparation.

E.11 SAMTRA (SAM TRanscription Analysis)

VERSION: 1.0, June 1991

HARDWARE / OS REQUIREMENTS: Standard Sesam workstation Description:

- The software package SAMTRA normalises and verifies transcriptions in SAMPA notation and computes phoneme and diphoneme distribution statistics. The data flow between the three modules of SAMTRA is via the intermediate files SAMTRA.NOR and SAM-TRA.STA (which are accessible for further developments). The modules perform the following actions:
 - 1. Module 1: Transcription verification
 - 2. Module 2: Computation of phoneme and diphoneme distribution
 - 3. Module 3: Flexible printing of statistics
- Procedure:
 - 1. SAMTRA reads two input files, a SAMPA symbol file and a SAMPA transcription file, and tries to write a normalised transcription to the file SAMPTA.NOR.
 - 2. If there are errors in the transcription file, SAMTRA prints them out on the screen and to the file SAMTRA.NOR.
 - 3. If the normalisation procedure succeeds, SAMTRA reads the normalised transcription file SAMTRA.NOR and computes the phoneme and diphoneme distributions, and writes them to SAMTRA.STA.
- The text output of SAMTRA is dependent on the file SAMTRA.SET (which overrides the default printout constants) and command line switches.

DEVELOPING LAB:

Faculty of Linguistics and Literary Studies Universität Bielefeld P 100131 D - 33501 Bielefeld Tel: (49) 521 106 3510 Fax: (49) 521 106 6008 Email: gibbon@spectrum.uni-bielefeld.de

Contact: Dafydd Gibbon

E.12 SOAP (Speech Output Assessment Package)

VERSION: 4.0, March 1992

Hardware / OS requirements: Standard Sesam workstation, with AU21/22 Oros board

DESCRIPTION:

- SOAP is a PC-based system for outputting speech and collecting the responses of a listener.
- The protocol (in terms of audio/visual cues, the file/path/part of speech files, time-out error strategy, etc.) can be defined by the user.
- The listener can enter responses via the keyboard or using a mouse/screen interface with a user-defined screen display.
- The software comes with demonstrations of SAM standard assessment protocols for testing at four levels: segmental, suprasegmental/SUS, prosodic and quality.
- Various utilities allow automatic randomisation of test sequences, the interpretation and scoring of test responses and automatic rulebased generation of segmental test files.

DEVELOPING LAB:

Dept of Phonetics and Linguistics 7406 University College London Gower Street London WC1E 6BT Tel: (+44) 71 387 Fax: (+44) 71 383 0752

CONTACT: Kate Jones

E.13 PTM (Parametric Test Manager)

VERSION: 1.0, March 1994

HARDWARE / OS REQUIREMENTS: Extended Sesam workstation:

(Windows operating system and a MCI-compatible sound card, 800x600 pixel graphics system). There are no special memory or hard disk requirements. The hard disk capacity depends on the amount of speech files which the user wants to use. The PTM software needs 4.5Mb for installation where 2.4Mb are used by the sample test files.

DESCRIPTION:

The new Speech Output software (SO software) of the SAM-A Project was designed to provide a tool which covers most aspects of speech assessment. This tool is called Parametric Test Manager (PTM). The PTM system consists of different databases and programs. Hence PTM is not a single application, but a software package.

The PTM includes the following features: (1) resource data management, (2) test management modules and (3) demonstration test. The resource data management includes the facilities to allocate stimuli, their acoustical realisations and subject data. From these data test corpora or subject groups can be created.

The new Speech-Output software does not only deal with data representation, but also with execution of a test. The software provides facilities to support people who want to carry out a certain test. A test execution can be divided into three phases: (1) test definition and preparation, (2) stimulus presentation and (3) scoring and documentation. The PT design includes applications to perform typical tasks as data acquisition, test list generation and session supervision.

DEVELOPING LAB:

Ruhr-Universität Bochum (RUB), GERMANY Lehrstuhl für allg. Elektrotechnik und Akustik Ruhr-Universität Bochum 44780 Bochum GERMANY Tel: (+49) 234 700 6652 Fax: (+49) 234 7094 165 Email: hegehof@aea.ruhr-uni-bochum.de

Contact: Thomas Hegehofer

SAM (1992). "Summary of SAM Software", in: ESPRIT Project 2589, Multi-lingual speech input/output assessment, methodology and standardisation. Final report, year three, 1 III 91 – 28 II 1992, Ref: SAM-UCL-G004, pp. 37–46.

F EUROPEC recording tool

EUROPEC : [European Program d'Enregistrement de Corpus]

 ${\tt License}\ :\ {\tt non-commercial}$

Author : J. Zeiliger

Task :

- Record and playback
- Protocol-driven recording session.
- Visual and acoustic prompts.
- Start and endpointing of utterances
- Signal and label files production

DESCRIPTION : This is a corpus prompting/recording software making easier the recording of speech databases. It ensures constant recording conditions for various speakers and/or corpus. The recording session is driven by a protocol file (user-made) which describes the way items (components of the corpus) will be displayed on the screen.

 $COMPONENTS \ :$

- "Standard SAM format" files production.
- One or two-channels recording and listening facilities.
- Three recording modes (manual, automatic, continuous).
- Parameter adjustment (sampling rate, input and output gain, etc.).
- Sound levelmeter.
- Free / operator controlled session.
- Error treatment.
- Acoustic prompt.
- Easy data verification.
- Safeguarding files.
- Easy data retrieval and manipulation.
- Simplified session.
- User's Guide and many example files.
- User-customisable in any language.

Modularity :

EXTENSIBILITY :

- Size : Version 4.11: 170 kbytes.
- (data: examples of speaker, corpus, protocol and label files provided.)
- Implementation : Microsoft C (v5.1)

TMS code for the I/O board

PLATFORM : SESAM workstation (PC-DOS based), with OROS AU21/22 I/O board. LANGUAGES : All. RETARGETABILITY : ORTHOGRAPHY : Labelling via ASCII. EXAMPLES : The program is used in laboratories all over Europe. STATUS :

- robust (used in over 30 labs).
- stable
- Windows release under consideration.

DOCUMENTATION : User's guide (30 Pages) SAM-ICP-O45, March 31st, 1991 UPGRADES : No foreseen upgrade. Windows implementation under consideration. SOURCECODE : Provided to members of ESPRIT projects only CONSULTING : FORMAT : 1.4 Mb HD floppy disk PRICE : One floppy and mailing costs.

RESTRICTIONS :

- Research purpose only.
- Citation

 $\operatorname{Contact}$:

Address : Institut de la Communication Parlée, Grenoble, France.

The EUROPEC prompting/recording tool has been clearly designed to make easier the recordings in large recording campaigns for the constitution of speech databases. However, it provides also simple and basic functionalities. Some particular points must be developed here:

DATA AUTHENTIFICATION

The EUROPEC software ensures automatic and consistant filenaming of the files it produces. The uniqueness of a filename during a recording campaign is ensured by an incremental serial number. For each recording session, the components: speaker, corpus to be recorded, recording conditions, protocol to be used, have to be registered in the relevant DB files.

PROTOCOL OF CORPUS

The protocol-based approach uses a protocol file that contains instructions to be executed by the program during the recording session. These instructions address the prompting of items on the screen and the design of the session. They define how and when items will be displayed, and allow training phases before the recordings. Here are some examples of the relevant codes implemented:

- CLS: CLearScreen and put cursor to 0,0
- VON: Video attribute ON: center, blink, inverse, underline, bright ...
- VOF: Video attribute OFF
- BEL: ring bell
- MSG: message to display according to valid video attributes
- NWL: newline on screen

- PAU: n (integer) = duration of the pause in seconds
- TRN: set training mode, no recording
- CMT: comment
- GET: n (integer) n: number of items to read in the corpus text file, then to display and record (if recording mode is activated) or not (if training mode is activated)
- LOP: -n, m loop in the protocol file
- n: number of one-line-back in the present file m: number of iterations
- REC: recording mode activation

In a research environment a prompting/recording tool can easily be extended or retargeted with useful functionalities to cover other experiments and researchers needs. For instance, local variants of the EUROPEC software have been successfully produced to perform adequate experiments in the following domains:

- PERCEPTION: The acoustic prompt can be used to perform perception tests. Stimuli are played to the listener that is then presented menu options to indicate his answer. The answers are stored in the label file (results of the test).
- PROSODY: Online listening facility is provided when the recordings aim at matching a particular ('spontaneous') prosodic profile. The session is a sequence of record, listen/accept/reject (menu options), record, listen/.. phases.
- Acoustic: The acoustic prompt can be used to play blank noise during articulatory positioning then simultaneous I/O is used to record the production of the vocal tract excitated by a blank noise.

All these examples demonstrate that prompting, recording and playback facilities can appear in different ways in various experiments. The protocolbased approach as initiated in EUROPEC is adapted to the definition of the experiment session by the user. This approach must be clearly enhanced: a recording tool should be better considered as a software platform providing basic functionalities as signal input/output, acoustic or text prompts, storage of labels or results, mouse driven menu. The modularity of these functions would allow the user to design his experiment and write the corresponding protocol file in a high level language; the protocol being then interpreted by the program all along the session.

G Digital storage media

This appendix lists the available storage media for digital data and summarises their capacity, accessibility, and longevity in a table. The technical data given here reflects the state of the art in mid '95. A personal estimate of workstation configurations for speech processing for the near future (1997) is given in the bottom row of the summary table.

G.1 Storage media dimensions

The basic dimensions along which storage media can be described are

- permanence, and
- order of access.

Other characteristics, e.g. capacity, speed of access, or cost, are also useful in describing storage media.

G.1.1 Permanence

In *volatile* memory, the memory contents are lost when the power supply is interrupted. Typically, the main memory of a computer is volatile. Data is held in integrated circuits (IC) and is refreshed periodically by supplying the circuits with current pulses.

In non-volatile memory the contents are retained even without the supply of power. Non-volatile storage media are special read-only (ROM) integrated circuits, magnetical, magneto/optical, or optical media. ROM chips are programmed by applying a strong current to the individual memory locations which burn a specific pattern into the chip (there also exist programmable ROM chips whose contents can be erased through ultraviolet light and reprogrammed with strong currents). In magnetic media, the magnetic stratum is polarised by a magnetising current, and this polarisation remains after the current has been switched off. In magneto-optical media, the data itself is held in the magnetic stratum; the optical tracks are used for a precise control of the read and write head. In optical media, different degrees of reflectivity of a stratum are exploited to store data. Data is written to the medium either by burning or by mechanically pressing a pattern into the optical stratum.

G.1.2 Order of access

In random access each memory location can be accessed independently of any prior memory access. Main memory, also called random access memory (RAM), floppy disks, hard disks, and CD-ROMs are typical random access media.

In *sequential access*, access to a memory location depends on prior memory accesses. Typical sequential media are tapes. A memory location on tape can only be accessed after having read the entire tape before the memory location.

G.1.3 Other storage media characteristics

The most common storage media characteristics are

- CAPACITY is given in byte (one byte = 8 bit = one octet). Note that the terms kilo-, mega-, giga-, and terabyte are used ambiguously because they can be based on the decimal (kilo = 1000) or the binary system (kilo = $1024 = 2^{10}$).
- SPEED OF ACCESS is measured in seconds and refers to the time needed to access a given memory location. In mechanical systems, e.g. disk or CD-ROM, the access time is largely determined by the *seek time*, i.e. the time needed to access a specified track and find a particular memory location.
- DATA RATE is given in bit or byte per second and refers to the amount of user data that can be transferred within a period of time. *Raw* data rate refers to the total amount of data transferred, i.e. administrative data exchange with the device controller plus the user data. In general, the raw data rate is much higher than the user data rate.
- COST of storage is determined by the cost of the storage device, the storage medium (for interchangeable media like tapes, hard disk cartridges, etc.), and the cost of maintenance. The cost of medium can be given in absolute or relative terms, e.g. cost per megabyte.
- LOCATION of a storage device is either *internal*, i.e. built into a computer, or *external*, i.e. a separate device. Internal devices occupy slots in the bus and require that there is sufficient space and power supply. External devices attach to the computer via peripheral connectors, e.g. SCSI-ports or PCMIA slots, and usually they have their own power supply. In principle both external and internal devices can be addressed through the same protocol, e.g. SCSI; however, different interfaces are very common, depending on the quality of the hardware and the bus architecture.

G.1.4 Compression

Sampled data files can be compressed and by doing this one can win on average 50% of the storage space. The amount of possible compression depends on the signal-to-noise ratio in the recordings. A high signal-to-noise ratio results in a better compression.

Compression of telephone speech (A-law or U-law) results in a compression of about 40%, as was experienced during the storage of the Dutch POLYPHONE material. A compression tool that is often used and is recommended here is *Shorten* by T. Robinson. It is available via anonymous ftp from svr-ftp.eng.cam.ac.uk.

G.2 CD-ROM

Today, CD-ROM (or CD for short) is the most widespread storage medium for large amounts of data due to the

- HIGH CAPACITY: A CD can hold up to 650 megabyte of user data. The data rate of current ('95) quadruple speed CD-ROM drives is approx. 600 KByte/s of user data.
- LOW PRICE: The cost of a pressed CD in larger quantities is less than 2 \$, and CD drives are inexpensive.
- STANDARDISATION: The physical structure of a CD is defined in a small number of industry standards, and is independent of the logical organisation of the data.
- LONGEVITY: Pressed CDs may hold data without loss for 40 years and longer.
- ROBUSTNESS: CDs are virtually unbreakable, tolerate small scratches or stains on the medium, and can be cleaned easily.
- VISUAL IMPACT: Copyright notices, title and author information, and decorative labels can be applied to the medium directly.

CD drives can either hold a single CD or have a CD changer subsystem which holds up to 7 CDs (changing CDs in a multi-CD system requires about 5 secs).

The major restriction of CDs is that they are a read-only storage medium. A CD is mastered on a hard disk the contents of which are then copied to a tape. This tape is sent to a production site which produces a *glass master* CD which then serves to produce metal master CDs. These metal master CDs are used to press the final CD copies.

CD-R is a write once/read many (*WORM*) storage medium. In contrast to pressed CDs the data is recorded on special virgin CD-R disks by burning a pattern into the reflective substrate of the CD-R. This process is irreversible. A CD-R can be read by any CD-ROM drive.

Producing CDs individually has two important consequences:

- CD-R contents can be customised to user demands, and
- CD-Rs can be given unique serial numbers, e.g. to track unauthorised distribution.

The major drawback of CD-R is that their longevity is reduced. CD-Rs come with a one year limited warranty only. However, one can expect data to be safe on CD-R for approx. 10 years if the medium is not exposed to high temperatures or ultraviolet radiation (which may affect the reflective stratum).

G.2.1 Physical CD formats

Philips and Sony as the inventors of the CD have defined the physical structure of CD in a small set of standards, the (*red, yellow, green, and orange books*). Two of these standards are of relevance for data CDs:

- 776 Part IV: Spoken language reference materials
- CD-ROM DA is the original audio CD for up to 74 minutes of 16 bit audio signals sampled at 44.1 kHz. This format is also suited for other types of data. It is described in the red book.
- CD-XA is a multisession format in which data can be written to the CD in a sequence of sessions. CD-XA is described in the yellow book.

The CD-XA format has now almost completely replaced the DA format for the storage of data, even if data is written in one session only. The green and orange books are for CD-i and CD-R respectively.

G.2.2 Logical data organisation

The logical data organisation on CDs is independent of the underlying physical structure. As a consequence, there exist CD implementations of all major operating system file systems, e.g. UNIX, DOS, and Macintosh.

Furthermore, a common file system has been defined in ISO 9660 (also called the *High Sierra format*), which is supported by all major operating systems. It follows the DOS file system structure closely by allowing only the legal DOS file names, i.e. eight capital letters plus a three capital letter extension (and an optional version number). The directory structure may be eight levels deep. An important extension of the ISO 9660 format is the Rock Ridge Extension, which allows symbolic links and nesting deeper than 8 levels.

Finally, due to the independence of physical structure and logical data organisation, hybrid CDs can contain different file systems in different *tracks* on one CD. This reduces the amount of space available to each file system, but results in a CD that has a native file system for more than one operating system.

G.2.3 CD-R software

There exist various software packages for the mastering and burning of CD-R for all major platforms. Until recently, however, substantial technical expertise was necessary to produce CD-R.

On the PC and the Macintosh, CD-R mastering and burning software has become available that takes care of many of the low-level details automatically. Basically, CD-R software creates either a *virtual* or a *real* (many CD-R applications call this *physical*, although this is not correct) image of the CD-R to be produced. This image is *not* a bit-for-bit copy of the final CD-R, it only organises the filesystem to be copied to the CD-R so that it is identical to the final filesystem on the CD-R. In a virtual image, only the pathnames of the files to be put on CD-R are stored, so that the source files can remain in their original location. Virtual images are compact, so that multiple virtual images can be held on disk simultaneously, and files need not be stored redundantly which helps avoid having inconsistent file versions. The major drawback of virtual images is that the lookup of files via pathnames may not be fast enough for the burning process, especially with double or quadruple speed CD-R drives. A real image contains the source file (or filesystem) itself. Each real image thus requires free hard disk space at least as large as the contents of the final CD-R, and files are duplicated which may lead to inconsistent file versions.

True bit-for-bit copies of CDs cannot be held on hard disks because of the different physical structure of CD and hard disk. However, some CD-R applications map the physical contents of a CD to a single file on hard disk, allowing true 1:1 copies of almost any CD-format to be made.

Note: A very good provider of CD-R soft- and hardware documentation, demo versions of software, and even public domain or shareware CD-R software can be found at the CD Archive at http://www.cdarchive.com/.

G.2.4 CD-R hardware

A CD-R recording station consists of a very large and very fast hard disk, a high-performance workstation, a CD-R drive, and an optional tape, CD, or removable hard disk drive.

The burning process may not be interrupted – otherwise, the CD-R in production is lost. A very fast hard disk provides source data at a sustained data rate high enough for the CD-R software. This data rate must be reliably achieved by the peripheral I/O components of the workstation – here, the raw power of the processor is less important than the sustained throughput of the peripheral I/O components, e.g. SCSI port.

Hardware for burning CD-R is becoming affordable -in 1995 CD-R recorder prices start at \$ 1000. For small volume productions, CD-R are a near-ideal solution: each CD-R costs about 10 \$ and can be produced individually in about 15 to 60 minutes.

CD-R recorders come in different flavours. Current CD-R drives feature a large buffer (at least 2 megabyte), run at double or quadruple speed, and support multiple physical and logical CD formats (at least CD-XA and hybrid plus ISO 9660, DOS and/or Macintosh HFS).

A tape, CD, or removable hard disk drive is useful for importing the source files to the CD-R station.

G.2.5 Future developments

The amount of data that can be stored on a CD is determined by the wavelength of the laser used to read the data. With shorter wavelength lasers pits can become smaller and can be packed tighter.

The first demonstration of prototypes of double sided high density CDs with capacities of approximately 2.5 gigabyte have been presented in 1995. They will reach the market in 1997.

Rewritable CDs are currently being developed, and prototype systems have

been presented already. However, the number of write access to such rewritable CDs is too limited to be of practical use. Also, it is not clear whether rewritable CDs have a future: hard disks are 50 times faster than CD drives, and in many cases the read-only restriction of CDs is absolutely required (e.g. for corpus distribution, copyrighted data, etc.).

Access time and data rate of CD drives will be improved by smaller and thus lighter read heads and higher spin rates of the CD-drives. Currently drives with a spin rate six times as high as in standard CD drives are entering the market, and further improvement can be expected.

G.3 Summary

Table G.1 summarises the characteristics of storage devices. It reflects the state of technology in mid '95.

	Memory	Disk					Tape			
	RAM	Floppy	Hard	Magneto Optical	CD	CD-R	QIC	Exa- byte	DAT	
volatility	volatile									
storage life ^{a}	-	2y	5y	10y	40y	2y	$_{2y}$			
technology	CMOS	mag	netic	mag/opt	0	ptical	magnetic		ic	
access		random						sequential		
mode	1	read/write many			read	read many/	read/v	vrite	many	
					only					
access time	10^{-9} s	$10^{-1} \text{ s} 10^{-3} \text{ s} 10^{-3} \text{ s} 10^{-1} \text{ s}$		0^{-1} s	t/2 ^b					
data rate	$10^9 \mathrm{b}^c$	10^2 b	10) ⁷ b	10 ⁶ b		10^6 b			
\cot^d										
$device^{e}$		\$500		500	\$200	>\$1000	\$600			
medium	\$50/M	1/M	.5/M	1/M	.01/M	.03/M	\$1/M \$0.02/M		2/M	
workstation										
$config^{f}$	16M	1.4M	$500{ m M}$	$500{ m M}$	$600\mathrm{M}$		$250{\rm M}$	200	0 M	
#		2	1	$\operatorname{external}$	1	external	ex	terna	l	
future (97)										
config	64M	1.4M	$4000 \mathrm{M}$		$1200\mathrm{M}$					
#		2	1	$\operatorname{external}$	1	external	external			

Table G.1: Storage media classification

 a y is years. Times given are manufacturers' warranties and claims respectively. For CD-Rs, most manufacturers offer only one-year warranties, which means that data has to be written to the CD-R within one year.

^bt is rewind time for the entire tape

 c b is bits/s

^dcost estimates in US\$ per device or Megabyte

^efor devices with interchangeable medium only

 $^f\!$ standard configuration of a high-end PC (Pentium or PowerPC processor) or middle range UNIX (Sparc) workstation

In the near future, workstations are expected to be equipped with the same number of devices as today (i.e. still one hard disk, one CD drive), but of larger size or higher speed. In general, an improvement of 2 to 8-fold can be expected. Optical media, including CD, will profit from shorter wavelength lasers currently being developed in research labs, but it will take until '97 for higher capacity devices to reach the market.

Tapes will continue to be used for archival and backup storage mainly due to the slow sequential access. Their capacity will increase -5 to 12 gigabyte ExaByte tapes have been presented. Some of the gain in capacity is the result of on-line compression.

G.4 References

Pohlmann, Ken (1992) The Compact Disk Handbook, A-R Editions Inc., Madison, Wisconsin, 1992

Red, Yellow and Orange book CD standards.

H Database Management Systems (DBMSs)

A Database Management System (DBMS) is a software system with the following properties:

- Data definition and manipulation in a formal data model,
- safe storage of large amounts of data,
- application independent storage of data,
- controlled access to the data.

In the past, many spoken language (SL) corpora have been stored in the file management system provided by the operating system. Data is accessed via shell scripts, custom programs, and application packages, e.g. spread-sheets, statistics programs, etc.

However, there are many disadvantages to this approach:

- Data is organised in an ad-hoc way.
- Data structures reflect the underlying physical data structures and the file system organisation.
- No automatic protection against data corruption.
- No access control to the data.

Clearly, this approach should not be followed.

H.1 Data model

A *data model* is a formal description of *entities* and their *relationships*, and of *operations* allowed on the entities. A data model is independent of a specific DBMS implementation. An *instance* of a data model or *database schema* is a data definition based on the data types and the data manipulation commands provided by the DBMS.

 $Data\ modelling$ is the process of describing the world of interest in the terms of a data model.

The following data models can be discerned:

- \cdot hierarchical
- $\bullet \ network$
- relational
- $\bullet \ {\rm object-oriented}$
- deductive

The development of the data models can be characterised as a) continued abstraction from physical storage, and b) increase in expressive power.

H.1.1 Hierarchical data model

In the hierarchical data model, entities (for instance, speakers, recordings, or types of speech) are considered as *record types*. Record types are subdivided into *fields*. A *record* stores the information of one particular entity.

Record types are organised in hierarchical *tree* structures. Except for the topmost record type (the *root*), each record type in the tree, a *node*, has exactly one predecessor, and zero or more successor record types. A node with zero successors is a *leaf*. A record belongs to exactly one record type, and it may be linked to zero or more records in the successor record types. The hierarchical data model thus considers all relationships between entities as 1:N relationships. Access to record types is possible by *navigation* through the tree, starting at the root; records are selected from the set of records of a record type.

H.1.1.1 Summary

- Hierarchical organisation of entities in 1:N relationships
- Access through navigation starting at the root record type
- Obsolete data model
- Implementations: IMS

H.1.2 Network data model

In the network data model record types are organised in a *directed graph* structure. A record type may have zero or more predecessor and successor record types. As in the hierarchical data model, the information on an entity is held in records. Unlike fields in the hierarchical data model, the fields of a record in the network data model may have multiple values.

A set type describes a 1:N relationship between two record types. Set types have names to distinguish the relationships of a record type. Since a record type may be linked to any other record type, complex relationships, e.g. N:M relationships, may be represented.

Access to record types is possible via special "entry point" record types from which dependent record types can be reached through navigation.

H.1.2.1 Summary

- Directed graph organisation of entities with named 1:N relationships between any two record types
- Access through navigation starting at special "entry point" record types
- Obsolete data model
- Implementations: UDS, DBMS-10, IMF

H.1.3 Relational data model

The relational model is a data model with a thorough mathematical foundation. An *attribute* is a named set of atomic values, the *domain*. A *relation*

table is a subset of the cross-product of the relation attributes.

The relational data model separates the logical data definition from physical storage.

Two data manipulation languages have been developed for the relational data model:

• *Relational algebra* is a procedural language which consists of the set operators union, intersection, difference and the relational operators selection, projection, and join. The selection operator selects rows from a relation table that meet a selection condition. The projection operator selects specific attributes from a relation table. The join operator merges relation tables according to a comparison condition over attributes from the original relation tables.

The result of a relational operation is again a relation, so that the operators can be nested.

• *Relational calculus* is a declarative language based on first-order predicate logic.

An expressionisа term of the form $\{x_1, \ldots, x_n | COND(x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m})\}$ with x_i variables over a domain and COND a formula that is either true or false. The variable bindings on the left side of the vertical bar are the values returned as the result of a query, and COND consists of atomic formulas which are connected through logical AND and OR operators.

The relational data model makes no assumptions about the relationships between entities. In general, an entity is mapped to a relation table, with the properties of an entity described through the relation attributes. Meaningful relationships between entities are expressed through relational algebra operations or calculus expressions. Hence, complex relationships between entities can be expressed, with the notable expression of recursive relationships.

SQL

Relational algebra and calculus are formal data manipulation languages. As such they lack data definition constructs, and because of their formal nature, they are not well suited for interactive data manipulation. The language SQL (*Structured Query Language*) is a relational database language which covers both data definition and manipulation and is close to English. It has become the de-facto language standard for relational DBMSs, and has been formally standardised by the ISO.

H.1.3.1 Example

The following example is a sample speaker and recording database schema implemented in SQL. First, relation tables are created using the data definition command CREATE with the appropriate arguments:

```
CREATE TABLE SPEAKER (

ID, CHAR(8), PRIMARY KEY

NAME, CHAR(20)

FNAME, CHAR(20)

SEX, CHAR(1)

DBIRTH, DATE)

CREATE TABLE RECORDING (

ID, DECIMAL, PRIMARY KEY

RECDATE, DATE

MEDIUM, CHAR(8)

LOCATION, CHAR(20)

SPK, CHAR(8)

FOREIGN KEY SPK REFERENCES SPEAKER)
```

The data types available in SQL are restricted to very simple character or number types of fixed length. Bit-stream data and complex data structures are not supported.

Then queries are formulated using the data manipulation language of SQL:

```
SELECT S.ID, S.NAME, S.FNAME, R.RECDATE
FROM SPEAKER S, RECORDING R
WHERE S.DBIRTH > 12/27/60 AND R.SPK = S.ID
```

Besides being a database language for interactive database access, SQL has become, through its standardisation, increasingly popular as a programming language interface. External applications generate SQL code, which is then transmitted to the DBMS and evaluated there. The result relation is returned to the calling application for further processing (either as a whole or one tuple after the other with a cursor mechanism).

Most commercial relational DBMSs support SQL, but add their own extensions to overcome the limitations of the standard data types, e.g. with binary large objects (BLOBs) for image, audio or video data, or complex data structures for graphics objects.

H.1.3.2 Summary

- The relational model is a mathematical model.
- Logical data types are completely independent of physical storage.
- Data is manipulated using procedural relational algebra or declarative relational calculus.
- Relational DBMSs are now standard DBMS technology.
- SQL is the de-facto standard relational database language, but each DBMS implementation extends the standard in a proprietary way.
- Implementations: DB2, ORACLE, SyBase, Ingres, and many others; available for all platforms (Mainframe, UNIX, PCs, Mac).

H.1.4 Object-oriented data model

The object-oriented data model aims at bridging the semantic gap between relation tables and entities of the real world through *objects* that directly correspond to entities. An object has a unique and immutable *object identifier*, and it belongs to a *class*. Classes are object definitions; they comprise attributes and operations over the class or the attributes. Both attributes and operations may be private to a class, i.e. visible only to objects belonging to that class (*encapsulation*), or public, i.e. visible to other classes. Classes are organised in class hierarchies and may inherit attributes and operations from superclasses. Operations are invoked by sending an object messages (*message passing*) which are executed if the required operation is defined for the object, otherwise they are passed on to the superclass.

In general, object-oriented DBMSs are integrated into an object-oriented programming language, giving them the full expressive power of the programming language and persistent object storage. Object-oriented DBMSs have just entered the marketplace, and they are successful in *non-standard* applications which require the full expressive power of programming languages, or complex and highly diverse data structures.

H.1.4.1 Summary

- The object-oriented data model is based upon the notion of object.
- An object belongs to a class, and classes are organised in a hierarchy with inheritance of properties from super to subclasses.
- Object-oriented DBMSs now enter the marketplace.
- There is not yet a standard object-oriented database language.
- Implementations: O₂, ObjectStore, POSTGRES, Starburst, GemStone and many others

H.1.5 Deductive data model

The deductive data model is a restricted first-order predicate logic extension of the relational data model. In the deductive data model, relations are either defined *extensionally* through *facts*, or *intensionally* through *rules*. Rules may be defined recursively.

Various formal languages have been developed for the deductive data model, e.g. Datalog⁺, a function-free sublanguage of first-order predicate logic, or Datalog which includes negation.

Logic programming languages, e.g. Prolog, extend Datalog with complex data structures. Current Prolog systems feature access to external DBMS or an internal database component for the persistent storage of large amounts of data.

H.1.5.1 Summary

- The deductive data model is a logic based extension of the relational data model.
- Data is stored as facts and rules.
- Persistent logic programming languages combine the declarative logic language with persistent data storage.
- Implementations: *LDL*, Eclipse, Prolog

H.2 Safe storage of data

DBMS provide safety features that prevent the loss of data due to hardware failure. The basic mechanism is that of *transactions*. A transaction is a sequence of data definition or manipulation commands that is considered as atomic by the DBMS; a transaction either succeeds completely, or fails and undoes all data changes (*rollback*).

During a transaction, the DBMS works on a copy of the data; if the transaction succeeds, the copy is written to permanent storage, otherwise it is simply discarded.

Transactions may be visible to the user. Other mechanisms to preserve the physical integrity of data are invisible:

- commit points, i.e. points in time at which the database contents is dumped to a permanent storage medium in a consistent way
- redundant data storage, e.g. through a *RAID*-array (Rapid Access to Inexpensive Drives) organisation of disks where redundant data is distributed over hard disks to allow the reconstruction of a corrupted hard disk, or *mirroring*, where exact duplicates of devices are used.

H.3 Application-independent storage of data

Data held in a DBMS has a greater lifetime than the applications using the data. In fact, in most cases data will outlive the DBMS in which it is stored. This means that data storage has to be organised independently of any application so that different applications (or generations of applications) can access the same data efficiently, and that import and export procedures have to be provided for the migration of data to new DBMSs.

In DBMSs, application independence is achieved by hiding the physical storage of data from the users. The DBMS automatically maps the storage requirements of a database schema to the file system of the operating system, creates a meta-schema which contains all information about the data stored in the DBMS and generates index files to speed up access to the data.

External applications cannot access data stored in a DBMS directly, e.g. by opening a file. On the one hand, this is a further security measure by which tampering with data is prevented, on the other hand this places the burden

of determining efficient data storage organisation and optimising access to the data on the DBMS.

H.3.1 Controlled access to data

In general, DBMSs are multi-user systems, i.e. many users or applications may access the same data in parallel. In order to prevent the interference of user operations, access to the data is controlled by the DBMS through user-specific views and access privileges.

A user-specific view or *subschema* defines which entities and relationships are visible to the user. The entities and relationships visible to the user must not be identical to those of the global database schema.

Access privileges define what mode of access a user has to the data. Access privileges can be defined for the global database schema, the subschema, or individual entity sets (e.g. tuples in a relation table, objects in a class). The access mode is either *read* or *write*. In general, a user can grant his access privileges to other users.

H.4 Summary

In SLP, DBMSs should be used for speech corpora of any size.

The special requirements of SLP, namely

- the size of signal data,
- the complex data structures for an adequate representation of data on different levels of abstraction, and
- the permanence of the data $% \left({{{\mathbf{x}}_{i}}} \right)$

cannot be met with the hierarchical and network data models.

Even the relational data model in its pure form is not well suited for SL corpora because it lacks support for bit-stream data and complex data structures. However, all current commercial relational DBMS implementations support binary large objects and a richer type system, which makes them candidate DBMSs for SL corpora. The major advantage of relational DBMSs is that they are now a standard and proven technology.

Object-oriented DBMSs, especially in conjunction with object-oriented programming languages, and persistent logic programming languages provide both a rich inventory of data types, full computational power, and a good support of the basic DBMS functionality. They are thus well suited for SL corpora.

The major drawback of object-oriented DBMSs is that there is not yet a common standard object-oriented database language. The major drawback of persistent logic programming languages is that there exist only a few

implementations of such systems. However, the logic based formalism is basically the same as in phonology and linguistics, and thus it is possible to use a single formalism on all levels from phonetics to linguistics.

The major advantages of a DBMS can thus be summarised as follows:

- Logically coherent data definition in a formal data model,
- uniform access to data in a database schema,
- stable database schema, flexible amount of data,
- interactive and application program interfaces for ad-hoc querying and inter-application communication.

DBMSs are, for instance, applied in the GRECO project (cf. Carre (1992), or the Alvey STA project described in Thomas & Winski (1987)), or the PHONDAT-VERBMOBIL database of spoken German (cf. Draxler (1995)). It will be clear that only the basic characteristics of DBMSs are mentioned above. The interested reader should consult Elmasri & Navathe (1989), or Ceri, Gottlob & Tanca (1990) or any other recent database introductory text book.

H.5 References

Carre, R. (1992) Speech databases. In: Ainsworth, W.A. (ed.) Advances in speech, hearing and language processing: A research annual. Vol. 2. London: Jai Press. 199–215.

Ceri, St., Gottlob, G., Tanca, L. (1990) Logic programming and databases, Berlin: Springer Verlag.

Draxler, C. (1995) An introduction to the PHONDAT-VERBMOBIL Database of Spoken German; Practical Applications of Prolog 95, Paris, 1995.

Elmasri, R. & Navathe, S.B. (1989) Fundamentals of database systems. Redwood City, California: The Benjamin/Cummings Publishing Company, Inc.

Thomas, T.J. & Winski, R. (1987) Speech recognizers, assessment in the laboratory, not in the field. *Speech Technology* 3(4): 88–93.

I Speech standards

I.1 Introduction

The development of speech technology is dependent on the availability of data and tools operating in various computer environments. Speech as acoustic data is digitised to be computable, and therefore the speech research community pays tribute to the various sectors concerned with the digitisation of sounds. It has to deal with the requirements of its purpose and with the available current technology.

Among these sectors, the audio industry takes over the hi-fi market with the development of Compact Disc technology and Digital Audio Tape. The computer industry was interested first in audio output, then, with the development of multimedia capacities, both hardware and software (operating systems, I/O boards) were affected to handle speech sampling progressively, then speech synthesis and recognition. Last but not least telephone and communication networking play an important role in the transmission and coding of speech and sounds as their worldwide scope implies of course standard protocols.

Each of these sectors has its own requirements and develop too many standards with regard to hardware and software aspects. A digitised signal is characterised by many parameters which can lead to various encodings on communication channels and to numerous file formats when it is recorded on a computable medium.

I.2 Signal theory recalls

When digitising signals, the following parameters have to be considered:

- sampling rate (number of samples per second),
- number of bits per sample (currently 8, or 16),
- digitisation law (linear, log, U-law, A-Law, etc.),
- number of channels (1 for mono, 2 for stereo, etc.).

These settings are those of the A/D converter when the data are recorded and must be used when they are played. They vary along the uses of different communities. Usually, the recorded signal is stored in files. When there are N channels, samples may respectively pertain to different files or sometimes to a unique one: in such a case samples or frames of samples are interleaved. The bytes order for each sample in the file (little or big endian) is also a relevant factor in reading the files (for example HI-LO order may be reversed from DOS to some UNIX environment). Data compression whenever used, implies the inverse uncompressing scheme to be applied before using the data.

I.3 Speech research

Inside the speech research community itself, formats used for speech databases differ widely, depending on their purpose and the applications they are designed for. The major differences concern the file format and the sampling frequency.

SAMPLING FREQUENCY:

There is a variety of sampling frequencies used in existing speech databases. Table I.1 shows a few examples:

Table I.1: Sampling frequencies

Sampling rate	Speech databases
8000 Hz	Telephone DB as POLYPHONE, CRICUBE (Canada)
$10000~{ m Hz}$	Read texts corpus (Netherlands)
$12000~{ m Hz}$	ATR DB (Japan),
$12800~{ m Hz}$	Telephone DB as COLLECT (Italy)
$16000~{ m Hz}$	ATR DB, ASJ (Japan), BDSONS (France), GRONINGEN
	(Netherlands),
	PhonDat & Erba (Germany), Timit (USA)
$20000~{ m Hz}$	ATR DB (Japan), EUROM-1 (Europe), BDBRUIT (France)
$48000~{ m Hz}$	Jeida (Japan)

The higher the sampling rate, the more space-consuming is the corresponding file, the same amount of acoustic data given. Ten seconds of speech sampled at 10 kHz correspond to a 200 kb file length (with a 16 bits quantification), but to an 800 kb file length when sampled at 40 kHz. If usual frequencies tend to be higher because of the high-quality technology being available and of the storage disk space becoming cheaper, many fields in speech research stick to middle frequencies for various reasons: because they do not need such a high quality for their purpose (for instance, speech synthesis), because they are linked to technology standards (European telephony standard: PCM A-Law (POLYPHONE)), or because their purpose is to deal with low-quality speech (speech recognition) for real applications.

FILE FORMAT:

It is very unlikely that, getting a speech file, one can guess its sampling rate, coding, conditions of recording and, even less, the age of the speaker. So it is of crucial importance that information on the speech signal file must be somehow available in order to use it properly. The minimum information required concerns of course the way of accessing the file (byte order, quantification, sampling rate). But information about the recording con-

ditions, the speaker characteristics, the text of the utterances, and various parameters, is more than useful in real speech studies. There are two main philosophies in force: storing information *within* the speech file or *outside* the speech file, i.e. in an external file. These two approaches have both pros and cons, and are well-represented respectively by the NIST/SPHERE format and the SAM format.

1. NIST/SPHERE:

This format is provided by the National Institute for Standard and Technology in the USA, and makes use of a *within* approach using a SPHERE header. It consists of an "object-oriented, 1024-byte blocked, ASCII structure which is prepended to the waveform data. The header is composed of a fixed-format portion followed by an object-oriented variable portion." "The fixed portion is as follows:

NIST_1A 1024

The first line specifies the header type and the second line specifies the header length." The remaining object-oriented variable portion is composed of object-type-value "triple" lines which have the following format:

<TRIPLE> ::= <OBJECT><space><TYPE><space><VALUE><OPT-SPACES>

The currently defined objects cover database identification and version, utterance identification, channel count, samples count, sampling rate, min and max level, and A/D settings. "The list may be expanded for future databases, since the grammar does not impose any limit on the number of objects. The file is simply a repository for "standard" object definitions. The single object "end_head" marks the end of the active header and the remaining unused header space is undefined" (but within the 1024 bytes limit)".

The NIST/SPHERE format is widely used in US and elsewhere, also for US & Dutch POLYPHONE and French BREF. It is supported by NIST, maintenance path exists, and it is provided with a set of tools to handle the header (access, update, remove, replace ...). The header approach minimises the risk of losing track of data identity; the header can support both prompt and transliteration texts but requires data files to be changed after collection for annotation and also if an upgrade/correction is issued. The header is fixed-length and unpromptable through a text editor.

2. SAM:

This format is a European 'standard', defined by the SAM consortium (ESPRIT Project "SAM": Speech Assessment Methods) (see Appendix C). SAM claims for an *outside* approach (headerless) using an associated

description file. It consists of a speech file + associated description file. A speech file contains only speech waveforms. An associated description file (ASCII) is linked to the speech file.

The files come in pairs; their names are identical, except for the last letter of the extension, according to SAM terminology. The associated description file is a standard label file with a header and a body. It contains all information which is usually required by people working on the files without the database management system. A label file is made of a header and a body (or several). Each line consists of a specific mnemonic followed by the corresponding value:

<MNEMONIC> : <space><value> (optional <comma><space><value>...)

The	label	heade	er	from	LHD :	
			to	LBD:		
the	label	body		from	LBD :	
			to	ELF:		
or	a new 1	label	body	from	LBD:	
			to	ELF:		

In a current annotation file the header contains database identification, file localisation, file production, A/D settings, sampling rate, start and end samples, number of channels, speaker information, and pointers to the prompt textfile, recording conditions and protocol. As the format is potentially adapted to store several items in a file, the body contains on-the-field labels for the one or several items recorded in the speech file: sequence beginning (in sample), sequence end, input gain on recording, minimum sample value, maximum sample value, orthographic text prompt are present for each item. Discontinuities between the items are indicated if any. Both the content of the header and of the body can be extended to store new relevant descriptors or labels, provided that adequate mnemonics are created and no contradiction occurs with existing ones.

The SAM format is widely used in Europe for multilingual databases (EUROM-1) and for national ones (French BDSONS, English SCRIBE, Italian COLLECT, Spanish ALBAYZIN). The current SPEECHDAT consortium adopted the SAM format for its telephone recordings. (SAM provided a conversion routine from NIST to SAM format on the DARPA/TIMIT CD-ROMs). The associated description file implies to have files going together by pair and increases the risk of losing files. But the headerless system keeps data files unchanged after collection during database transcription

correction/upgrade. It supports multiple annotation levels. File length is not limited, an information is available through a single text editor. ELRA (European Language Resources Association) should take care of the maintenance/upgrades of this format.

3. Other format

The VERBMOBIL project in Germany has developed its own format, especially for handling dialogue. Examples of databases in Japan (such as JEIDA, ASJ) have no header.

DISCUSSION

So far we have seen that a correct description of a sound data file includes a lot of mandatory fields. The first (and minimal) one contains information on how to use the file:

- file format: header/no-header (and its length);
- number of significant bits, bytes order, sampling rate;
- type of encoding: linear (signed/unsigned), logarithmically, compression scheme, etc.;
- channel number: mono/stereo, etc.

But researchers in their job need a lot more about the speech they study:

- Information about the corpus: name, prompt text;
- information about the speaker: sex, age, accent, etc.;
- information about recording conditions: quiet room, office, car, noisy environment, etc.

The development of speech applications in new domains implies many other descriptors being available. Descriptions of new data types (multisensor, multimodal, dialogue) are needed, as well as more complex and complete descriptions of data (dialogue, e.g. in WOZ techniques; multimodal synchronisation; timing notations; additional descriptions such as dialogue flow, emotional state, man-machine situation). Furthermore, the forthcoming development of database distribution and networking will require information about the sources of the data to be available, such as the way of obtaining it and the right to use it.

The standardisation carried out in previous large collaborative projects must be clearly enhanced; efforts must be devoted to the representation of more complex information on speech data, with associated description files and pointers to various descriptors (including location of the data, source of the data, transformations applied to the sources, country of provenance acknowledgement, restrictions on use, derived information ...).

I.4 Computer hardware and software

Current hardware uses a variety of in-house or more widespread standards. For example, coding is 8 or 16 bits at Apple (Mac) and PCs, U-LAW 8 or 16 bits at SUN (Sparc), NEXT, VAX, DEC, U-LAW or A-LAW at HP. Available sampling rates are often limited to 8 kHz in the UNIX world, but higher rates may be available in the PC world (DOS/Windows) and Mac depending on the current or professional I/O boards.

File formats are often indicated with the filename extension they bear. Computer manufacturers such as NEXT and SUN deal with .au (AU) or .snd (SND) files, Apple and Silicon Graphics with .aif (AIF); I/O boards manufacturers may promote their own format (as .voc for SoundBlaster boards) and the developer of the Windows operating system, MICROSOFT, tries to impose its .wav (WAVE) format. This situation is complicated by the encoding mean (linear, compressed, data and information intermingled, etc.) and even for the same filename extension, the implementation may vary slightly for different operating systems (WAVE in Windows or UNIX environments, SND in NEXT or PC/Mac environments). A standardisation initiative comes through the development of Internet, promoting an interchange format called MIME.

A major example of the constraints imposed on the speech research community by the market can be demonstrated by looking at the implications of the multimedia standard development in the PC world.

Multimedia standard

The world of PCs has considerably evolved during the past few years along two relevant dimensions:

- Operating system: the Windows operating system is now used worldwide, and it provides a suitable graphics interface.
- I/O boards: the development of multimedia functionalities implied the availability of low-cost I/O boards to be easily included in a low-level PC configuration (SoundBlaster, Pro Audio SPectrum ...).

The point is now whether these current boards, primarily dedicated to audio output, can satisfy the needs of speech research and applications in terms of:

- signal quality (signal-to-noise ratio ...);
- sampling frequency: the multimedia standard is basically derived from CD Audio standard (44.1 kHz) or DAT one (48 kHz). So most of the multimedia compatible I/O boards use sampling rates that are provided through successive entire divisors of this basic frequency (22.5, 11.25 kHz, etc.). But sampling rates used in our current speech databases are at present 16 kHz, 20 kHz, ... Care should be taken that a continuum of sampling

frequencies could be available (lets say from 5 to 50 kHz) on these boards, to satisfy the requirements of the speech research community. It is foreseen that all current cheap boards will not be convenient. Otherwise, on-line resampling techniques would be required (*) to maintain compatibility with existing databases, and for future databases the speech community is to adopt a standard "audio" sampling rate.

- file format: the multimedia standards apply the same. Most of the boards use "standard" (or peculiar) file format definitions, for example the main one is WAVE format (.wav). It means that these boards are not able to play the files from our existing databases (SAM or national) which are in a 16 bit linear format, as the WAVE one consists of chunks of data intermingled with chunks of encoding info. The files of these databases would have to be converted from one format to the other in order to be played. Future databases should either adopt a new "market" standard, or have their files converted on input and output.
- number of channels available (two or more channels may be requested for various microphones or sensor data recordings).

(*)(**) Using I/O boards without DSP implies that some signal processing will be deported to the PC (speech level detection, min/max measurement, eventual over- or undersampling). These on-line procedures, augmented with on-line format conversion routines, could increase the CPU load in such a way that low-level SESAM workstations could not be able to support running with a high speech sampling rate for example (or using two channels).

One topic is *background compatibility* with existing databases, another one is which format is going to be "the standard", i.e. the worldwide audio/ computer/speech standard. Such a topic is to be considered during the SPEECHDAT project, but it is foreseen that no unique standard will emerge and that conversion routines will remain a big issue. Many tools are available but as an example, even for the RIFF WAVE format the conversion between Windows and UNIX worlds is all but trivial. At the moment, it is not sure whether the current inter-changeable standard I/O boards in the market will satisfactorily meet the speech research needs or not, depending on the target application.

I.5 Conclusion

At this time of large-scale development of linguistic resources, their wide dissemination and re-usability, the problem of the standardisation of tools and data is especially relevant in the speech domain. The useability of speech databases and the portability of speech tools is very dependent on the functionalities provided or imposed by hardware and software manufacturers. The audio industry particularly (with the high sampling frequency used in DAT, CD) and computer manufacturers with their devoted I/O boards or sound systems formats (WAV, AU, SND files ...) do not make things easier. Although some powerful conversion tools do exist (as "sox" or "sfconvert"), there is a general lack of reliable and easy-to-use conversion routines. A close relationship between industrials and public research organisations should take care of the bridges among their respective standards. If no agreement on a common standard can be aimed at, at least import and export drivers for conversion to the different formats must be handled.

I.6 Appendix: Useful anonymous ftp sites AUDIO FILES FORMATS:

- ftp.cwi.nl:/pub/audio (part 1 and part 2), or
- rtfm.mit.edu:/pub/usenet/news.answers audio-fmts/part[12]

FILE CONVERSIONS (SOX: "Sound Exchange")

• ftp.cwi.nl:/pub/audio/SOX<version>.tar.Z (UNIX) and sox5dos.zip (PC)

NIST-SPHERE

- jaguar.ncsl.nist.gov in compressed tar form as "sphere-v.tar.Z"
- ftp.cwi.nl:/pub/audio/NIST-SPHERE (short description)

RIFF WAVE (.WAV)

- ftp.cwi.nl:/pub/audio/RIFF-format
- ftp.microsoft.com:/MSDN/CD8/RIFFNE.ZIP (complete description)
- teeri.ouli.fi:/pub/msdos/programming/*
- ftp.ircam.dr:/pub/music

J EUROM-1 database overview

The first SAM database, EUROM-0, was the precursor to the much more substantial EUROM-1 recording undertaken in SAM. This database has been widely distributed on a single CD-ROM and contains five hours of speech material recorded with 16 kHz 16 bit sampling using a condenser microphone in anechoic rooms from four single accent speakers in each of five languages (English, French, Dutch, Italian and Danish). NATO single and triple digit sequences are recorded using only the speech signal, and a continuous speech passage, with a common numeric theme across languages, is also recorded here using two channels – with both speech and laryngo-graphic inputs. This CD-ROM has been used extensively in the SAM project and is specified as a reference for calibration of the speech input assessment tools.

A subsequent major data collection activity has resulted in the collection of a very substantial amount of data which is unique in the size and the breadth of its coverage of different European languages. The EUROM-1 database contains more than twelve hours of data for each of the eleven European languages covered: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish. The material is of high acoustic quality, and was selected specifically for use in the assessment of speech technology devices.

The control software used in making the recordings provided for orthographic labelling of the data and alignment of the text and signal portions at the level of the prompt units. Phonotypical transcriptions have been made separately for all languages and broad phonetic labelling using SAMPA (SAM Phonetic Alphabet) has been applied to some parts of the database. Language subsets of EUROM-1 are now available on CD-ROM, and the provision of EU funding is planned to ensure availability of all recorded material of this important reference resource. A minimum of three CD-ROMs are planned for each language.

Very careful consideration was given to the homogeneity of the data across languages. This was achieved by the use of identical recording protocols, which were specified earlier in the project and applied using standard software tools, and by a careful definition of the speech content, such that each language was represented in the same way wherever possible. The speech (and calibration) recordings were made in acoustically treated rooms using calibrated condenser microphones and, in addition to the acoustic signal, larynx activity was recorded simultaneously, using a laryngograph, for samples of the speech in each language. The use of anechoic condenser microphone recordings permit the subsequent imposition of post-production effects. The recordings were made using the SAM agreed standard of 20 kHz 16 bit sampling to ensure optimal signal representation; inter-utterance acoustic background signals were also preserved. The protocols defined for collection of database materials have been developed to provide guidelines on recording procedures and quality criteria for use in the wider European Speech Community. These are provided in full in Appendices B and C. The database material specification for each language is as follows:

- C(C)VC(V) material in isolation and in context, in the range 60–100 items per language
- 100 selected numbers from 0-9999, providing complete coverage of the phonotactic possibilities of the language number system
- 40 short passages comprised of five the matically linked sentences
- 50 sentences composed to compensate for phonemic frequency imbalance in the passages
- 5 pairs of context words for use with C(C)VC(V) material

The database has been designed with a hierarchical structure to maximise its usefulness both for training and testing different types of speech technology device and for more basic research including inter-language comparisons. In each language, material was recorded by 60 subjects, 30 female and 30 male, each of whom recorded 100 numbers, 3 passages and 5 sentences. Of these a "few talker" subset of 5 females and 5 males made extended recordings: isolated C(C)VC(V) items, 500 numbers, 15 passages, 25 sentences. A further "very few talker" subset of one female and one male, selected from the 10, additionally recorded the contextualised C(C)VC(V)s and the 5 context words, using both acoustic and laryngographic signals. A total of 660 speakers thus recorded over 130 hours of data, making this a very substantial multilingual resource with many different applications.

K Polyphone project overview

This paper¹ gives a short description of the creation of the Dutch POLY-PHONE corpus. It then proceeds to show how that corpus has already been put to use in the development of a number of practical applications and how it is used for research purposes.

Applications that are touched upon include a fully automatic Train Time Table Information system and the automation of collect call and phone card calls.

The paper concludes with remarks on and recommendations for future corpus development.

K.1 Introduction

From a technological and algorithmic point of view speech recognition in the Public Switched Telephone Network (PSTN) has reached a quality that is sufficient for a range of commercially attractive applications, at least in the U.S.A. and Japan, where large populations use the same language and recognisers can be trained for the native language. Although the recognition technology is basically language independent, similar applications in most European countries lag behind Japan and the U.S. This technology gap has no simple explanation; it is due to a complex of factors, some of which are cultural in nature, while others are organisational or financial.

It is probably true that Europe has a cultural attitude that is less open towards new technologies than the U.S.A. and Japan. However, all European language areas have less inhabitants than either the U.S.A. or Japan. In terms of native speakers German is the biggest European language, with some 100 million speakers, if one is willing to include the German part of Switzerland and Austria, even if the varieties of the language spoken in these countries differ from standard German at least as much as American and British English. Compared with 245 million Americans who speak English and 125 million Japanese, it is evident that return on investment in speech technology will be smaller in Europe.

Number of speakers is not the only factor determining traffic volume for a given application; cultural factors also play a role. The number of calls placed from residential areas in the U.S.A. is substantially higher than in most European countries, because the telephone is used in a different way. Even within Europe the differences are large. In the Netherlands the transition from local calls to a single national premium rate number for public transport information caused virtually no protests from the public, whereas

¹T.I. Boogaart¹& L. Bos²& L. Boves^{1,2,3} ¹PTT Research, ²SPEX, ³Nijmegen University Paper presented at IVTTA 1994 Japan.

the German Railway Company feels that such an action would cause an uproar in their country. Perhaps the fact that the transition was accompanied by a substantial improvement of the service (more information – also on busses and trams – and more accurate information and shorter waiting times) has helped to keep protests down. But the fact that premium rate services are becoming widely accepted in the Netherlands has also been of importance. Moreover, toll free numbers as a weapon in the battle for the consumer plays virtually no role in Europe. Thus, the fact that a certain application has proved to be a commercial success in the U.S.A. cannot be simply generalised to the European market.

In a situation where the commercial viability of speech technology applications is not always evident, cost and ease of application development become automatically of paramount importance. Managers lose interest in new technologies that require substantial capital investments if the return is insecure. They also quickly lose interest if application development takes longer than a few months.

K.2 Reusable resources

Both investment capital and development time can be reduced substantially if reusable tools and resources become available. Speech corpora to develop and train recognition systems are the most expensive and time consuming resource needed to build applications that use speech recognition in the PSTN. Moreover, annotated speech corpora are also indispensable for running objective performance tests of off-the-shelf technology that has proven adequate performance in some languages, but not yet in the language of the country of interest. Because large commercial interests are at stake with the introduction of automated services using ASR, one can understand that marketing managers are reluctant to accept new technologies as long as there is no definite proof of their adequacy for the home market.

In the fringe of ICSLP-'92, the international Cooperating Committee on Speech Databases and Assessment (COCOSDA) defined guidelines for a corpus that should be able to solve at least some of the problems slowing down the uptake of speech technology applications. Comparable corpora, now known under the name POLYPHONE, should be recorded for as many different languages as possible. In the course of 1993 these guidelines were made more concrete by the Linguistic Data Consortium (LDC), which specified a recording protocol for American English and Spanish [1,2]. Taking those protocols as a point of departure, PTT Research and the Speech Processing Expertise Centre SPEX set out to record the Dutch POLYPHONE corpus.
K.3 The Dutch POLYPHONE corpus

K.3.1 Recording workstation

The recording workstation used for POLYPHONE was based on an Aculab telephone interface, a Rhetorex Voice Card and driver software, Show-'n-Tel application development software, and a 16 port operational license, in an OS/2 PC. Each item recorded was stored in a separate file; all files were copied to a UNIX network for transliteration and archiving.

The recording platform was set up to record and store the speech signals in A-law format. Because the Dutch PSTN is completely digital, the acoustic quality of the recordings is determined by the characteristics of the caller's local loop and the background noise in the caller's location.

K.3.2 Speaker selection

Prospective callers received a personalised letter. Originally, we aimed at collecting 5000 speakers, uniformly divided over a large number of cells, defined according to four criteria, viz. (1) geographical region, (2) socioe-conomic status, (3) sex, and (4) age. It should be emphasised that the uniform sampling of the cells was mainly motivated by scientific arguments: in order to find the funds for creating the corpus it was necessary to make it attractive for a wide range of linguistic research, including sociolinguistics and dialectology. Perhaps part of the speakers in our corpus will not be heavy users of the automated services that can be developed by means of the Dutch POLYPHONE corpus. However, we trust that a wide coverage of language and speech behavior will lead to applications that are more robust than what could have been obtained with recognisers trained with much more restricted speech material.

Geographical region, operationalised as the province in which the speaker lives, is the best practically feasible approximation to regional accent and dialect background. By sampling provinces, we sidestep the unsolved problems of how many different regional accents should be distinguished and how these should be defined. Due to the very uneven distribution of the population over provinces it appeared to be practically impossible to get equal numbers of speakers from each province [3,4].

Socioeconomic status is difficult to define, and even more difficult to assess reliably from what respondents are willing to say. We decided to approximate status on the basis of the education level of the respondents. We distinguished three levels, viz. (1) only primary school, (2) secondary school and (3) college/university. Using hindsight, this division was somewhat unfortunate: in formal terms almost every person younger than about 60 has been to school until at least the age of 16, so only a very small proportion of the population falls into the first category. Thus, it is not surprising that we were able to recruit very few speakers who said that they had no more than elementary school. The numbers in the remaining two classes are approximately equal.

We distinguish four age classes, i.e., under 20, between 21 and 40, between 41 and 60, and 61 and older. Information about age is acquired by asking the respondents for their year of birth. Since we set a minimum age of 16 for participation, the under 20 group is much smaller than the other groups. The group of 61 and older is also underrepresented. The group between 20 and 40 is about 50% larger than the group between 40 and 60.

K.3.3 The speech material

The speech material recorded in the POLYPHONE project consists of 32 read items, 14 extemporaneous answers to printed questions, and 4 extemporaneous answers to questions not printed on the response sheet. The material *to be read* consists of the following items:

• 12 numbers (digit strings, natural numbers, guilder amounts, postal codes)

- 2 city names
- 4 application words (e.g. start, accounts payable, tomorrow, etc.)
- 3 spelled words, designed in such a way that all letters of the alphabet occur approximately with equal frequency
- + 1 date and 1 time
- 4 sentences with an application word
- 5 phonetically rich sentences, designed in such a way that each speaker would realise all phonemes of Dutch at least once.

The following list of *printed* questions is asked:

- Is Dutch your native language?
- Did you ever live in another country than the Netherlands?
- Would you volunteer to participate in another study like this one?
- What is your last name?
- What is your house number?
- What is the name of the street you are living?
- What is your postal code?
- In which city do you live?
- In which cities did you grow up?
- Are you a man or a woman?
- What is your age?
- Which code represents your education level?
- (1 = primary school, 2 = high school, 3 = college/university)
- Please, say a familiar phone number.
- Please, give your comments about this recording session.

The following *unprinted* questions are asked:

- Please, spell your name.
- Are you calling from your home phone?
- Are you using a cordless telephone?
- What time is it now?

K.3.4 Postprocessing

Postprocessing was done at the Dept. of Language & Speech, Nijmegen University, using software running on a PC under MS-Windows, equipped with a Pro-Audio board. Whenever the answer is predictable (i.e. in all cases where the caller is supposed to read preprinted material), the expected answer is displayed on the screen.

Postprocessing consists of four steps, viz. (1) word-by-word transliteration of all items, (2) transliteration of extra sounds and noises, (3) collecting demographic data, (4) assessing the quality of all items. The students who carried out the work were instructed to do the tasks in exactly this order. On average, a recording session took slightly less than 20 minutes to process. We have seen that the sentences contain substantially more dysfluencies than the other items.

K.4 Use of POLYPHONE in application development

At present, a number of application development projects are under way in the Netherlands that employ ASR techniques. In this section we describe the way in which POLYPHONE helped to enable these projects.

K.4.1 Train time table information

In a collaborative project between PTT Research, Philips Research Aachen and the Netherlands Organization for Scientific Research NWO we are working on the development of a Dutch version of the Train Time Table Information service that is already available for German and that is described in another paper in these Proceedings. The system we have in mind can best be characterised as a guided mixed-initiative dialogue system: the system will ask specific questions, like *From where to where do you want to travel?*, but it will allow the user to give under- and overinformative answers. If an answer is underinformative, the system will ask explicit questions to elicit the missing information. When an answer is overinformative, for instance when the caller adds the desired arrival time to departure and destination station, the system will try to process that additional information too. If time and date information is not offered spontaneously, again the system will ask explicit questions to obtain it.

Clearly, a Train Time Table Information System is an application intended to be used by the general public. Moreover, most users will call the system only occasionally, so that one cannot rely on users getting acquainted with the peculiarities of the service. Pilot experiments carried out by the *Nederlandse Spoorwegen*, the Dutch Railway Company, have shown that the part of the public who need Time Table information are not able nor willing to deal with a menu-based interface. Thus, there seems to be no alternative for starting an automated service with a mixed-initiative dialogue system. To implement a Dutch version of the Train Time Table Information System a number of steps must be taken:

- Train a phoneme based continuous speech recogniser for Dutch, designed to handle telephone speech.
- Design the phonemic recognition lexicon for the application.
- Design a language model suitable for use in this application.

The POLYPHONE corpus has been instrumental in all these steps.

K.4.1.1 Training of the recogniser

For the development of the phoneme based recogniser use has been made of the phonetically rich sentences in POLYPHONE. Following the approach that has proved successful for German we have started with a recogniser based on context independent phone models. The recogniser has been trained assuming that the automatic grapheme-to-phoneme transcription of the transliteration data is correct. That assumption is probably wrong to some extent: Dutch has quite some pronunciation variation at the phonemic level. At the time of this writing we are using the POLYPHONE recordings for an empirical investigation of the range of that variation. Up to now, researchers had to be content with rather subjective ideas about this crucial issue.

K.4.1.2 Building a phonemic lexicon

An essential part of the recognition engine in a Train Time Table system is a lexicon comprising phonemic representations of station names. Here too, there is non-negligible pronunciation variation. Since all station names have been read by at least five speakers, we can use the POLYPHONE recordings to make an inventory of the pronunciations. This is especially relevant for the names of the smaller stations, since pronunciation variants for larger stations can also be collected by other means.

K.4.1.3 A model for yes/no expressions

Virtually every information dialogue contains yes/no questions. In previous applications of ASR in telephone information systems for the general public it has appeared that there is quite some variation in the way people answer these questions. Since POLYPHONE contains four yes/no questions, all to be answered spontaneously, we have a substantial amount of data to build a model of the answers.

The analysis of the answers that we have performed so far confirms the existence of substantial variation; yet, it appears that the very large majority of the expressions adhere to a simple schema, so that it is easy to build a model. We have seen a large difference between the two items for which we expected affirmative responses: Almost 93% of the subjects used a single word (e.g. *ja*, *jawel*, *jazeker*) to confirm the assumption that Dutch was

their native language; the proportion of one word confirmations dropped to 75% for the question whether the caller was willing to participate in another recording session. Very few callers said "no", but the way in which they expressed their confirmation was much more varied.

83% of the subjects used a single word (e.g. *nee*, *neen*) to deny that they ever lived abroad for an extended period of time. Most of the people who used more complicated expressions did so to tell us in what foreign countries they had lived. 80% of the callers used a single word to deny that they were using a cordless phone; over 13% of the callers said they were using a cordless phone.

A detailed analysis of the more verbose answers showed that only a very small proportion of the affirmative answers contained *no*-words and that the same is true for negative answers and *yes*-words.

Another observation worth mentioning is that politeness forms like *yes*, *sir*; *no ma'am* were virtually absent. This may be due to the fact that the *yes/no* questions were located in the last part of the recording session, when the callers should be fully aware that they were talking to a computer. However, it is also possible that what we see reflects the growing casualness in the Dutch society, where "speaking with two words" is quickly becoming the exception rather than the rule.

All these observations confirm our expectation that the NLP module in our system should be able to handle the large majority of the *yes/no* expressions that will be used by the callers. In confirmation subdialogues in an information system (e.g. after the caller has given departure and/or destination station) the language model expects an affirmative expression, but negations may occur due to errors of the recogniser. The POLYPHONE corpus contains a number of examples of negations where confirmations were expected. We are working on a closer analysis of these cases, to find out whether they contain systematic syntactic structures that could help in making the language model more specific.

K.4.1.4 Time and date expressions

In previous experiments with information and reservation systems it has appeared that – quite surprisingly – linguists do not have accurate models of the way in which people express times and dates. The POLYPHONE corpus contains a large number of these expressions. Currently we are analysing the syntax of these expressions in order to build models for use in the Train Time Table Information system. Unfortunately, the expressions used by the POLYPHONE speakers are to a large extent determined by the way in which the items were printed on the response sheets. No spontaneous expressions of dates or times were obtained. This will make it very difficult to derive reliable estimates of the relative frequency with which individual expressions will occur.

K.4.2 Phone card services

PTT Telecom has started a campaign to promote Operator Services, like Collect Call and Phone Card calls. The PTT Telecom Phone Card is marketed under the name of *Scopecard*. As in other countries, operator services are expensive, mainly because of the high costs of the personnel. Thus, it is obvious that Telecom is looking for ways in which these services can be automated.

One simple way of automating Phone Card calls is to connect customers with an IVR platform that handles the recognition of the card number, the PIN code and the number to be dialled via DTMF; this is how a large proportion of domestic calls are automated. However, few operators offer the capability to connect toll-free lines to an IVR application in another country. Also, in many of the countries where PTT Telecom customers spend their vacation rotary dial phones still form the large majority. Thus, large scale automation of card services implies the deployment of ASR. The same goes for automatic collect calls, of course.

K.4.2.1 Automating card services

For applications like automated card services it is not enough to have a recogniser that can handle isolated or connected digits *in abstracto*. For the real-world performance of the application a solid knowledge of the way in which customers pronounce card and phone numbers is at least as important, since that knowledge can be exploited in designing application specific language models. POLYPHONE has provided us with a rich source of information about the way in which Dutch people express these numbers.

Two items in POLYPHONE that are related to telephone numbers were analysed. The first pertains to numbers read from the response sheet. All these numbers were printed in the same format, i.e. area code, dash, subscriber number (e.g. 020-5252183). The second item consists of answers to the question *Please, say a familiar telephone number*. In discussing the results we will use the term *digit* for the words *zero, one, ..., nine*; the term *number* will denote numbers between 10 and 99.

Presently, the Dutch PTT's number plan has two groups of area codes, one comprising three digits (like 020 in the example above) and one comprising five digits (e.g. 08894). Subscriber numbers can have four to seven digits. Because transliteration does not include intonation markers, it is not possible to discriminate between three and five digit area codes. We doubt whether the transliterators would have been able to parse all answers to the request to give a familiar number correctly.

The format of the read numbers is quite different from the format of spontaneously produced familiar numbers: in read numbers the proportion of *digits* is much larger than in spontaneously pronounced numbers. It is also

worth mentioning that 18% of the read and 23% of the spontaneous numbers contain extra sounds, far more often preceding the number than following it.

POLYPHONE provides similar information about the way in which long card numbers and shorter PINs are pronounced. At the time of this writing these expressions are still under analysis.

K.4.3 Speaker verification

Advanced telephone applications will inevitably grow from pure information systems to mixed information and transaction systems. Security and fraud prevention then become major issues. Card Services of Dutch PTT is investigating the possibility of using speaker verification as one means of fraud combatement. In its original specification the POLYPHONE corpus is not suited for research into speaker recognition. Nijmegen University, in collaboration with the Dutch National Forensic Science Laboratory, have made additional recordings of 100 speakers who have called eight times, using different handsets. Half of the speakers is recruited from the Nijmegen area, the other half from the Hague, in order to minimise dialect differences. Also, the speakers form 50 pairs of brothers or father—son, in order to allow us to investigate whether speaker recognition techniques can be fooled by close relatives.

K.5 Requirements for future corpora

At this point in time it is too early to give very specific recommendations for future speech corpora. Yet, a number of things can be said. In future corpus projects an attempt should be made to obtain spontaneous expressions of time and date, in addition to the read expressions in POLYPHONE.

Speaker selection and recruitment is still a difficult issue. In the Dutch POLYPHONE project much time, effort and money was spent in order to get a maximally uniform sampling of a large number of cells. To a considerable extent, these efforts have been to no big avail. The major reason to strive towards uniform sampling was scientific: we wanted the corpus to be as attractive as possible for linguists and dialectologists, of course without interfering with the requirements of speech technology. The latter requirements are ill-defined. It is quite likely that applications like Train Time Table Information must deal with the public at large, including low income groups whose speech may differ from the general standard. More research is needed to clarify this issue.

This research was supported by the Foundation for Speech Technology, which is funded by the Dutch National Program for the Advancement of Information Technology (SPIN).

References

[1] Godfrey, J., Graff, D. & Martin, A. (1994) "Public databases for speaker recognition and verification." *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification*,, Martigny, 5–7 April 1994, pp. 39–42.

[2] Bernstein, J., Taussig, K. & Godfrey, J. (1994) "Macrophone: An American English Telephone Speech Corpus for the POLYPHONE Project." *Proc. ICASSP-'94*, Adelaide 19–22 April 1994, pp. I–81–I–83.

[3] Boves, L., Boogaart, T. & Bos, L. (1994) "Design and recording of large data bases for speaker verification and identification." *Proc. ESCA Workshop Automatic Speaker Recognition, Identification and Verification*, Martigny, 5–7 April 1994, pp. 43–46.

[4] Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. & Boves, L. (1994) "Creation and analysis of the Dutch POLYPHONE Corpus." *Proc. ICSLP-'94, paper 30.5.*

L European speech resources

CAVEAT:

The present report has been achieved under the account of the RELATOR Project(LRE-62056). It is the fruit of the work of the speech partners involved in this project. They have tried to accurately represent the existing resources in Europe, after having contacted colleagues and compiled various sources of information. But they are sure that there are emission and errors in this status report and they apologise in advance and welcome corrections/additions/comments. An extended version of the survey is available as an annex of the RELATOR WP1 deliverable.

L.1 Overview

Spoken language is central to human communication and has significant links to both national identity and individual existence. With the increase in availability and capabilities of computing resources, there has been and will continue to be a large expansion in computer-based language technologies. These technologies include speech recognition and synthesis, vocal access to information retrieval systems, speech understanding (or spoken language) systems and spoken language translation. Central to progess made in spoken language technologies lie large corpora of speech with associated text, transcriptions, and lexica.

The structure of spoken language is shaped by many factors, including the phonological, syntactic and prosodic structure of the language being spoken, the acoustic environment in which it is produced, and by the communication channel. The speech signal is produced differently by each speaker, each with a unique vocal tract which assigns its own signature to the signal. Speakers have different dialects, accents and speaking rates, and their speech patterns are influenced by their emotional and physical state, and the context in which they are speaking (e.g., reading aloud, in conversation, giving a lecture) and the acoustic environment. Due to the many sources of variability in the speech signal, a great deal of speech data are needed to model different speech characteristics, and in particular, different dialects and accents.

Recent activities, such as the creation of the Linguistic Data Consortium (LDC) and the Center for Spoken Language Understanding at the Oregon Graduate Institute $(OGI)^1$ in the U.S. and the LRE RELATOR project in Europe, national efforts in Japan, Australia and China, as well as the international Coordinating Committee for Speech Databases and Assessment

 $^{^1\,{\}rm OGI}$ has collected an 11-language telephone-based corpus which has been used for common evaluation of language identification algorithms. This corpus is currently available through the LDC.

(COCOSDA), point out the growing worldwide awareness of the need for and importance of large, publicly available common corpora for the development and evaluation of language technologies, particularly speech recognition and spoken language understanding, as well as for the development and assessment of speech synthesisers. These corpora allow scientists to study, understand, and model the different sources of variability, and to develop, evaluate and compare speech technologies on a common basis.

Corpora collection in Europe is the result of both national efforts and efforts sponsored by the European Community. Several ESPRIT projects have attempted to create comparable multilingual speech corpora in some or all of the official European languages. The first multilingual speech collection action in Europe was in 1989, consisting of comparable speech material recorded in five languages: Danish, Dutch, English, French, Italian. The entire corpus, now known as EUROM-0 includes 8 languages: Danish, Dutch, English, French, German, Italian, Norwegian, Swedish. Other corpora resulting from CEC projects include: SAM/SAM-A EUROM-1 (11 languages: Danish, Dutch, English, French, German, Greek, Itaian, Norwegian, Portuguese, Spanish, Swedish), ARS (Adverse Recognition System: Italian, English? **langs**), POLYGLOT (7 language IWSR database: Dutch, English, French, German, Greek, and Spanish, 5 language TTS database: Dutch, English, French, German, and Greek), ROARS (Robust Analytical Recognition System: Spanish, ?? **langs**), SPELL (Interactive System for Spoken European Language Training – French, Italian and English), SUNDIAL (Spoken language queries in the travel domain for English, French, Italian and German), SUNSTAR (Integration and Design of Speech Understanding Interfaces; English, German, Danish, Spanish and Italian), and ACCOR (cross-language acoustic-articulatory correlations: Catalan, English, French, German, Irish Gaelic, Italian and Swedish).

What follows is a brief status of linguistic resources for the European Countries, as well as a summary of some of the corpora resulting from European Community projects.

L.2 Criteria for assessment of the situation of Spoken Language Resources

As our main concern is language, the situation of spoken language resources is assessed on a per language basis. This does not correspond exactly to the situation on a per country basis, as several countries may contribute resources to the development of the same language (for example, Belgium, Switzerland and France all have French linguistic resources), or one country may contribute resources for several languages (Switzerland may contribute to French or German, and the UK has produced the Oxford Acoustic Phonetic database which contains spoken data for several European languages).

While ideally all the European languages should be adequately represented, it is clear that from a commercial standpoint the importance of a language depends on the potential market demand. While in the European market all European languages are of potential interest, large companies tend to prefer English, German, French, Italian and Spanish. This leaves an even stronger need for multilingual EEC initiatives in this domain which can pay attention to the under-represented European languages, as in at least the immediate future such needs cannot not be expected to be filled by industrial demand.

To aid in assessing the current situation in Europe, we provide guidelines on characterising the spoken language resources, and then summarise the main speech resources in Europe and identify the respective actors in their production.

L.2.1 Types and specificities of corpora

There are as many types of corpora as relevant factors which can be used to define them: speakers, texts, speech type, recording conditions, tasks and so on. Among this wide range of possible corpora, we may characterise them according to their intended use:

- 1. EXPERIMENTAL RESEARCH: These corpora are widely used for speech technology development and assessment. Much of the basic material was collected several years ago, and more recent technology requires more advanced materials.
 - 1.1 BASIC MATERIAL: Numbers, Words, Sentences, Logatoms
 - Number of speakers: medium (100-500)
 - Several repetitions.
 - 1.2 Advanced material: Continuous speech, passages, situated dialogue.
 - Number of speakers: small to medium (10-200)
 - Recently the trend has been to increase the number of speakers in such corpora
 - 1.3 Specific databases: multi-sensor corpora (Lx), articulatory, acoustic, video databases.
 - Number of speakers: small
 - These corpora tend to be relatively expensive to collect and may require sophisticated recording facilities and sensors, as well as specialised operators.
- 2. GENERAL-PURPOSE TELEPHONE CORPORA: These used for speech recognition and coding over the telephone. These type of corpora are relatively easy to obtain (the speaker only needs to call a specified telephone number) and relatively cost effective. However, with advances in communication technology, some of the problems currently posed by the limited bandwidth and noisy communication channel of today's telephones can be expected to disappear.

- Material: word lists, numbers, spelled names
- Number of speakers: large (several hundreds to several thousands)
- 3. APPLICATION-ORIENTED CORPORA: For specific tasks and/or environments (many of which involve the telephone network). By essence of their application specificity, many of these corpora are not easily reused for other applications.
 - Domains: Information retrieval travel inquiry (train and flight information and reservation), leisure activities, telephone services.
 Vocal dictation – medical, legal and insurance areas.
 - Adverse environments: Car database, Handicap applications.
 - Number of speakers: variable (small to medium)

The two extremes in corpus type are on one side very specific corpora for fundamental research, which may require complex recording conditions with multi-channel recordings, and a low number of speakers, and at the other application-specific corpora which may be recorded over the telephone with a large number of speakers. In addition to the recorded speech signal, we must highlight the importance and effort required to ensure that the appropriate associated information is provided. This associated information depends heavily on the type of corpus, but at a minimum must include revelant speaker information, transcriptions (at a minimum an orthographic transliteration), prompt material in the case of read-speech corpora, lexica, noise or channel characteristics and details of the recording configuration.

L.2.2 Actors in speech resource production

- 1. NATIONAL AND/OR ACADEMIC INITIATIVES: In every country where there is a history of speech research, academics (universities, speech research labs) produce the databases they need. However, these corpora tend to be specific to the needs of the producer and rest the property of the producer.
- 2. EEC INITIATIVES: EEC projects are a catalyst of production. Both for basic research and application-oriented databases, they are a way of developing links between academics and industry. Corpora also tend to remain the property of the consortium. As the conortium members are both academic/research and industrial, the needs cover all areas.
- 3. TELECOMMUNICATION/TELEPHONE SECTORS: Major telephone operators historically have been interested in speech technology, and most have their own research centers which collect the necessary corpora for their research activities. Most of these corpora are not publicly available.
- 4. INDUSTRY: Companies developing or integrating speech products need application-oriented databases. This is true both at national and international level, where foreign languages represent viable market opportunities. The data that can be provided by industry is varied, but for the most part unknown, other than that resulting from EEC initiatives.

The more of these actors are present in a given country, the more we tend to find a developed speech community both with existing linguistic resources,

but also with a strong demand for additional resources. As the speech community grows and the number of speech-based products extends, the amount of needed resources also grows.

a) need 1.1, 1.2, 1.3	provide 1.1, 1.2 with a), b), c) provide 1.3 with a)
b) needs variable	provide 1.1, 1.2 with a), d) provide 2 with c) provide 3 with d), a)
c) need 1.1, 2	provide 1.1 with a), b) provide 2 with b)
d) need 1.1, 1.2, 2, 3	provide 1), 2) with b)

So far, in reviewing the already existing resources, the presence of traditional actors, and the on-going projects, we assess the current situation as follows. Our starting consideration is that the under-represented European languages will need at minimum the resources that better-represented languages already have (at least the basic resources), and that well-represented languages will need still more resources. These needed resources will come from ongoing projects, and further needs can be foreseen through interviews with relevent actors in the speech research community and industry (ISC).

L.3 Summary of the current situation on a per language basis

L.3.1 DANISH (Denmark)

L.3.1.1 Existing databases and their actors

- Fundamental research: EEC initiatives produced basic/advanced material: EUROM-1, EUROM-0, SUNSTAR.
- Application-oriented: National initiative produced recorded/transcribed dialogues in the domain of flight ticket reservation and information.

L.3.1.2 Ongoing projects & new initiatives

- Experimental research: National initiative for advanced material (dialogue).
- Application-oriented: ONOMASTICA (EEC).
- Telephone: SPEECHDAT (Jydsk Telefon will collect a POLYPHONE type Danish Corpus).

L.3.1.3 Conclusion

There is a national academic effort. Telecom operator has a history of being involved in EEC projects and in the domain. Activities are carried out within National or EEC funded projects. The basic material exists (EUROM-1) but is still not available. No other known industrial initiatives.

L.3.2 DUTCH (The Netherlands)

L.3.2.1 Existing databases and their actors

- Experimental research
 - academics: GRONINGEN Corpus (basic research), BLOEMEN-DAL corpus, DEMSI corpus, Collections of speech recorded on analogue tapes. In coordination with the Reusable resources task group of ELSNET, the GRONINGEN corpus has been produced on CD-ROM for public distribution.
 - CEC: Polyglot, EUROM-0, EUROM-1, NOISE-ROM-0
- Application-oriented
- Telephone: Dutch POLYPHONE
- Lexica: CLEX (Dutch/German/English, distributed by SPEX and LDC)

L.3.2.2 Ongoing projects & new initiatives

- Experimental research: "Dutch National Corpus" (similar to BNC, 10 M words spoken material to be collected)
- Application-oriented: ONOMASTICA (EEC)
- Telephone
 - POLYPHONE: Dutch corpus of similar design, sponsored by SPEX and PTT Research.
 - Speechdat (EEC)
 - NWO Priority Program: information system about public transportation via the telephone.
 - EUROCOCOSDA, COCOSDA
 - Eurescom related Telecom cooperation.

L.3.2.3 Conclusion

The speech community in Holland is very active. Active participation in ESCA (president is L. Pols), ELSNET, EAGLES, EUROCOCOSDA, CO-COSDA. Strong cooperation between academics and telecom. Strong national effort. Industrial involvement unknown.

SPEX, the Dutch Speech Processing Expertise Centre located in Leidschendam, is interested to get involved, as the Dutch node, in a European infrastructure for database collection and exploitation. It distributes some corpora. EUROM-1 to be made available on CD-ROM.

L.3.3 ENGLISH (United Kingdom)

- L.3.3.1 Existing databases and their actors
 - $\bullet \ \, {\rm Experimental\ research}$

- 814 Part IV: Spoken language reference materials
 - Academics: SCRIBE, HCRC Map Task Corpus: British Normal Speech Corpus (advanced), Oxford Acoustic Phonetic database (8 languages).
 - EEC: EUROM-1, EUROM-0, Polyglot
 - Application-oriented
 - Telephone: Bramshill corpus of British English Home Office Telephone recordings: (12 CDs) held by the LDC.

L.3.3.2 Ongoing projects & new initiatives

English is represented in the EUROCOCOSDA (TED), EAGLES, and EL-SNET initiatives.

- Experimental research: National BE-WSJ0CAM, British National Corpus (BNC)
- Application-oriented: SQALE, ONOMASTICA (EEC)
- Telephone: Speechdat (Polyphone corpus to be collected by GEC Marconi)

L.3.3.3 Conclusion

Very active speech community in a variety of areas from fundamental research to technology development and applications. Many universities, companies, national telecom involved in the domain. Strong national effort, with active participation in EEC projects. Basic material seems to be available. EUROM-1 to be made available as multilingual resource (organised by A. Fourcin, UCL). However at the current time, the only resources whose dissemination are well-defined are those managed in cooperation with LDC. Other cases are by local arrangement.

L.3.4 FRENCH (France, Belgium, Switzerland)

L.3.4.1 Existing databases and their actors

- $\bullet \ \, {\rm Fundamental\ research}$
 - National initiatives produced basic/advanced material (BDSONS, BREF, BDBRUIT, ICY, SPOT)
 - EEC actions produced basic (EUROM-0, EUROM-1, POLYGLOT), specific (ACCOR) material.
- Application-oriented: There are indices of cooperation between academics (PSH/DISPE) and industry with the sponsorship of EEC (FREETEL, SPELL).
- telephone / Telecom: Speaker verification corpus (Switzerland)
- Lexicon: Bdlex

L.3.4.2 Ongoing projects & new initiatives

French is represented in the main European (and worldwide) current initiatives such as COCOSDA, EAGLES, EUROCOCOSDA, RELATOR, SQALE, ONOMASTICA, FREETEL and (through Switzerland) POLYPHONE, TED-Martigny. Some of these initiatives deal with linguistic resource standardisation, dissemination and production. The very next resources to be produced in French will be under the sponsorship of

- the French-speaking network: FRANCIL (Réseau Francophone de l'Ingénierie de la Langue): speech synthesis, vocal dictation, person-machine dialogue
- EEC: POLYPHONE type French corpus by Philips (SPEECHDAT), timetable enquires over the telephone (RAILTEL).

L.3.4.3 Conclusion

• Situation

The French speech community is quite active. All types of actors are present with many universities, academic institutions, and both public and private research laboratories are involved in national and EEC projects, in France but also in Switzerland and Belgium. The recent creation of FRANCIL demonstrates this activity. Major telephone, telecom companies operate in the field and the many companies demonstrate the interest in speech technology.

- Needs
 - telephone databases (Speechdat)
 - application-ready corpora (over-the-phone, spontaneous, different conditions, etc.)
 - advanced corpora (dialogue, dictation) (Aupelf)
- Dissemination

There are some attempt at making databases available on the market as commercial products, by the producer itself or through a licensed company. Most of the resources produced by academics are available free to other academics but distribution to industrials is still a problem and, when possible, is carried out on a case-by-case basis. As most academics have no way to handle properly commercial issues, the French organisation GRD-PRC is very interested in setting up a national repository that would take care of the dissemination of both written and spoken language resources. There is a strong desire for a European center for resource distribution, that would take care of issues such as IPR and licensing agreements.

L.3.5 GERMAN (Germany)

L.3.5.1 Existing databases and their actors

- Experimental research
 - Academics: basic (PHONDAT1), German Pronunciation Dictionary.
 - EEC: EUROM-1, EUROM-0, Polyglot
 - $\bullet \ {\rm Application-oriented}$
 - Academics: PHONDAT2, ERBA (train)

- 816 Part IV: Spoken language reference materials
 - Industry: Sunstar (EEC)
 - Telephone: Siemens Telephone Database

L.3.5.2 Ongoing projects & new initiatives

Various national, industrial, and EEC initiatives and cooperations, as:

- Experimental research
 - VERBMOBIL (basic, advanced) by industry and universities
 - VERBMOBIL-PHONDAT (advanced) by academics
 - TEDspeeches, TEDlaryngo (advanced), TEDphone by EUROCO-COSDA
 - Articulation of German Vowels (advanced)
 - Publicly Spoken German, very large database of german utterances
- Application-oriented
 - Onomastica (EEC)
 - Siemens "1000 read sentences" by company/university
- Telephone
 - TEDphone (EUROCOCOSDA)
 - SPEECHDAT (EEC) POLYPHONE type German Corpus by Siemens
 - Stemmer Telephone Database

L.3.5.3 Conclusion

- Situation: German is very active. All actors are present. Many universities, many large companies, the national telecom, are involved in the domain.
- Needs: Basic research material is more or less available, but making EU-ROMs available could be important as part of a really multilingual corpus. Application-oriented corpora are requested by industrials.
- Dissemination: Spoken language resources in German are either free for any use, available for research only, available for project partners only (EEC projects), or of unknown availability. So it is clear that in Germany, commercialisation of linguistic resources is an important issue.

L.3.6 GREEK (Greece)

L.3.6.1 Existing databases and their actors

- Experimental research: Basic material
 - EEC: EUROM-1, POLYGLOT
 - Academics: isolated speech 2000 words (alphadigits).

L.3.6.2 Ongoing projects & new initiatives

EEC: ONOMASTICA

L.3.6.3 Conclusion

Considering that Greece is a bit late for what concerns telephone equipment, speech technology is not the today crucial concern in Greece. So the investment of telephone companies is low, and there is no promising national initiative. The main basic research corpora have been produced due to EEC contracts, and it is therefore very important to achieve the work so far done (e.g. Basic material to be made available (EUROM-1)).

L.3.7 ITALIAN (Italy)

L.3.7.1 Existing databases and their actors

- Experimental research
 - EEC: EUROM-0, EUROM-1
 - National: IRST Acoustic Phonetic corpus of 3000 sentences, AIDA (CVCV words, digits).
- Application-oriented: ARS1000 database, SUNSTAR database (EEC)
- Telephone: Collect: telephone customers utterances

L.3.7.2 Ongoing projects & new initiatives

Italy is member of the COCOSDA, EUROCOCOSDA initiatives.

- Application-oriented: ONOMASTICA (EEC)
- Telephone
 - POLYPHONE: CSELT is collecting a large telephone speech corpus
 - including a complete POLYPHONE data set. - SPEECHDAT

L.3.7.3 Conclusion

- Situation: All actors present. Italian Speech community is active. Academics, telecom operator, companies do cooperate.
- Needs: Mainly telephone (SPEECHDAT) and application-oriented corpora. Advanced material for basic research is also needed.
- Dissemination: Basic research material (EUROM-1) was produced on CD-ROM and is being marketed by CSELT; the availability of other corpora is not well defined, but the general tendency is to produce the corpora on CDs with the possibility of being made available on a case-by-case exchange basis.

L.3.8 NORWEGIAN (Norway)

L.3.8.1 Existing databases and their actors

- Basic material: has been produced due to national (NTH) initiative and EEC initiative (EUROMs)
- Telephone corpora: are produced in cooperation by national telecom often in cooperation with Denmark.

L.3.8.2 Ongoing projects & new initiatives

- Experimental research
- Application-oriented
- Telephone
 - EEC
 - ONOMASTICA Norwegian Telecom, SINTEF
 - COST project 232 NTH multi-accented English speech cor-
 - pus collected over dialed-up international telephone lines
 - Cost project 249 NTH, Norwegian Telecom
 - National: Continuous speech recognition over the telephone line.

L.3.8.3 Conclusion

Research institutes, National Telecom, telephone operator are present. EUROM-1 to be made available as basic material. General lack of basic research corpora.

L.3.9 PORTUGUESE (Portugal)

L.3.9.1 Existing databases and their actors

- Fundamental research: EUROM-1 (CEC)
- Application-oriented: SUNSTAR (Portuguese accent for English words)

L.3.9.2 Ongoing projects & new initiatives

Portugal is represented in some European current initiatives such as ONO-MASTICA, RELATOR, ELSNET, SPEECHDAT and COST.

L.3.9.3 Conclusion

• Situation

The Portuguese speech community is relatively small, and its main actors have been so far Universities and academic research institutions. There is very limited funding for basic research in this area. Telecom companies have only very recently shown some interest in speech technology. There are extremely few databases for European Portuguese, although EUROM-1 will soon be made available as basic material.

- Needs
 - fundamental research (namely in terms of corpora with phonetic and prosodic labelling)
 - telephone databases (SPEECHDAT)
 - application-specific corpora
 - advanced corpora (dictation, dialogue)

L.3.10 SPANISH (Spain)

L.3.10.1 Existing databases and their actors

- Experimental research
 - EEC: EUROM-1
 - National
 - Corpus oral de referencia del espanol contemporaneo (Advanced) ALBAYZIN (Exp. and Appl.)
 - Automatic Speech Recognition PA85/86
 - Construccion de Sistemas de Reconocimiento de Habla Mediante Tecnicas De Aprendizaje Automatico TIC-448/89
 - Industry: TANGORA (IBM, also for other European languages (French, Italian??), but not available)
- Application-oriented: ROARS, POLYGLOT, SUNSTAR (EEC)
- Telephone

L.3.10.2 Ongoing projects & new initiatives

- Experimental research National: ALBAYZIN project, child language project (advanced)
- Application-oriented ONOMASTICA (EEC)
- Telephone Speechdat (EEC): Polyphone corpus to be collected by Vocalis.

L.3.10.3 Conclusion

All actors present. Many universities, research institutes public or privates, national telecom operator involved in the domain. Growing national effort, accompanied by a survey of textual and spoken corpora in Spanish. No corpus is currently available on CD-ROM. Availability and cost information not well-defined.

L.3.11 SWEDISH (Sweden)

L.3.11.1 Existing databases and their actors

- Experimental research
 - EEC initiatives produced basic material: EUROM-0, EUROM-1
 - National initiative produced basic material: Swedish sentence material
- Application-oriented National initiative produced CAR database.

L.3.11.2 Ongoing projects & new initiatives

- Experimental research
 - WAXHOLM dialogue project (national)
 - Onomastica (EEC)
 - A prosodic Swedish database project is under discussion.

L.3.11.3 Conclusion

Some research institutes are involved in speech technology. Effort for basic Swedish resources was supported the Swedish government to participate in EEC projects and by other national efforts. There is now national effort towards application-oriented and advanced material. But the basic material (EUROM-1) is still to be made available.

Status of telephone-based corpora unknown.

L.4 General conclusions

- Among existing and prospective resources, usability and reusability of speech databases produced through EEC projects should be enhanced. These resources are most of the time still unavailable several years after the end of the projects, because the time-intensive process leading from the raw data to marketable data was not initially included (or funded) in the projects (the marketable product is not a deliverable of the project). The distribution of these databases can be stimulated by the existence of an European Center for the Distribution of Language Resources who can provide funding for reorganisation and documentation of the already existing data and its production on CD-ROM.
- Existing databases whose diffusion is confidential should be highlighted. These databases are either not well-known because they remain in the scope of the project for which they have been designed, and/or have been kept confidential by their designers who do not want to disseminate them. It is obvious that for many industrials, linguistic resources are not yet considered as precompetitive resources, but as strategic ones. Any future European Center of LR production and distribution will have to manage the gap between the views and methods of industrials and academics: long-term widely available resources for academics vs. short-term strategic resources for industrials.
- There is a clear lack of speech databases suitable for multilingual evaluation. At the curent time, only EUROM has a Europe-wide dimension. Large, multilingual corpora are needed, and added value (full annotation and labelling) must be included in the design of new databases. It should be noted that the most widely used multilingual corpora are currently being recorded by OGI in the U.S. This corpus is being used for development and testing of automatic language identification techniques, which have a wide variety of practical applications.
- Pronouncing dictionaries and spoken language lexicons are very limited and should be developed for speech output systems assessment. Large phonetic lexicons are needed for use in speech recognition. One particular source that should be exploited is the outcome of the Onomastic project, and in particular the legal details for distribution of the lexica need to be worked out.
- A corpus of at least 100 hours of high quality speech should exist in each language providing coverage of many words and phonetic contexts, with

data from a reasonably large number of speakers (200-500). Most languages are far from that (for example at 20 kHz, 100 hours of speech correspond at 15 Gbytes, or about 25 to 30 CD-ROMs). In addition, this basic corpus will need to be supplemented by application-oriented corpora (such as for dictation (medical, legal, insurance domains), office systems, speaker verification/identification, topic spotting, information retrieval, etc.). Newspaper-based corpora are a potential multilingual source that should be encouraged, as these basic corpora provide at the same time speech and language modeling data, that can be used for dictation as well as other applications such as topic spotting. Multi-sensor and multimodal corpora are also needed for more basic research.

• For short-term commercial use there is a need for large telephone-based corpora, with speech from many (several thousands of) speakers covering a wide range of dialect, age, and socioeconomic backgrounds. These corpora should be extensive enough to permit the design of speaker-independent, vocabulary-independent speech recognisers that can serve as the source for a variety of applications.

Our conclusions are in agreement with those in the recent EAGLES report on Spoken Language Systems. It is clear that the need for adequate resources is a prime concern for many actors in the field of language technology. The following excerpts come from a draft EAGLES report on Spoken Language Systems (For EAGLES Restricted use). We highlight three areas of needs: speech corpora, lexicons and assessment of speech output.

- Speech corpora size (in terms of speakers), pp. 10–11
 - few speakers (< 5 speakers):for development of speech synthesis systems (dictionaries of phonetic elements). Advanced research. Multi-channel recordings (Electroglottogram, subglottal pressure, etc.)
 - medium (5–50 speakers): experimental research. Number of speakers and repetitions large enough for statistical processing / or for broad coverage of phenomena.
 - large (> 50 speakers): Train and test of speaker-independent recognition systems.

• Spoken language lexicons (p. 67)

"Large-scale spoken language lexical resources, from reference sources of standard, stylistic and regional pronunciations through vocabularies which are characteristic of spoken language, are required for current research and development with both statistical and knowledge-based technologies ... These spoken language lexical resources in the form of actual lexical databases and tools for constructing them, are sadly lacking." Assessment of speech output systems (p. 102)

"A short-term recommendation is to develop multilingual machine readable pronouncing dictionaries at the single word level which list permissible variations..."

L.5 Production costs

Speech database is the only type of speech resource we may be able to estimate the costs of production, based on concrete production cases. That is rough speech database. The cost of any Added Value (annotation, transcription, phonetic or prosodic labeling ...) will depend on the manual/semiautomatic/automatic way to proceed and on the corresponding tools available or to be developed. And this refer to Tools in general, whose development costs are those of the software industry (it is essentially a matter of specifications and corrresponding man-months). The production of spoken language lexicons and pronunciation dictionnaries should be comparable to speech databases in term of production costs. An estimate of this cost can hopefully be provided by the ongoing ONOMASTICA project, and perhaps by SPEX from their experience with CLEX.

A current estimation of the minimum cost for database production is 1 Ecu per utterance. But this is the lowest case which may be applicable to large telephone corpora, for which there is no need to move people (even not to pay them), and for which the initial equipment investment is relatively minor. However, these corpora are by nature of telephone speech quality and may not be useful for basic research and technology development. But many other factors can come to increase the cost. At the other extreme, multi-channel recordings in a specific and controled environment with representative selected speakers from all over a country, may be requested for advanced research material. The cost in this case, estimated from completed and available databases, may reach 10 Ecus per utterance (e.g. ten times more than for the telephone corpus.) In fact, more specific corpora, including a variety of articulatory sensors can be substantially more expensive. So we can argue that the cost scale for the type of widely applicable, general purpose speech data is from 1 to 10 Ecu per utterance, depending on various factors as:

- Type of utterance: word, sentence, passages, dialogue... the cost is higher and higher in terms of recording time, error recovery, storage capacity.
- Speech quality: sampling frequency and quantification (number of bits per sample) are relevant factors. From telephonometry to high quality standard audio files, the same utterance will require 5 times storage capacity (e.g. more CD-ROMs for the same database). The use of lossless compression techniques (T. Robinson, CUED) can reduce the needed storage by a factor of 2. The RELATOR project has supported the adaptation of the UNIX compression algorithm to DOS for use on PCs.
- Recording protocol: Multi-channel recordings require *n* times the storage space. For example, if both speech signal and Lx signal are recorded together, for one utterance the volume is doubled. The more the protocol is complex (multi-sensors, specific equipment) the more the production process is time-expensive and the pressing process is space- and hence time-consuming (number of CDs, effort to prepare them, etc.). Due to the

limited bandwidth of certain channels, the actual storage requirements may be less, as lower sampling rates can be used.

• Speakers selection: from people calling a toll-free number from their own phone, to speakers that must be present in a specific recording site (including transport, hotel, food expenses and eventual payments), there is a wide range of situations and corresponding costs.

M Transcription and documentation conventions for Speechdat

M.1 Introduction

In this appendix you find the transcription conventions for the SPEECHDAT corpora. The starting point of these conventions are the conventions used by LDC/ARPA in producing the ATIS CD-ROMs. The project has decided to simplify the transcription task to enable it to be performed quickly and to represent the most important acoustic events adequately for training and testing of automatic speech recognisers.

The SPEECHDAT corpora comprise 7 different languages: English, German, French, Spanish, Portuguese, Italian, and Danish. In the final section of this appendix some language-specific issues and choices are described. The documentation accompanying each language database will however describe fully all optional conventions and transcriptions used.

M.1.1 Points of departure

There are 6 points of departure:

- 1. The transcription is intended to be an orthographic, lexical transcription with a few details included that represent audible acoustic events (speech and nonspeech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance.
- 2. The transcription is intended to be a quick and broad transcription; transcribers should not have to agonise over decisions, but rather realise that their transcription is intended to be a rough guide that others may examine further for details.
- 3. Transcriptions should be made in two passes: one pass in which WORDS are transcribed, and a second in which the additional details are added. Background noises and uh's are easy to miss unless specifically attended to. It is recommended that transcribers have some background in phonetics and/or linguistics, or that their training and preparation for the transcription task cover some basics in acoustic phonetics and dialect and style variations.
- 4. The overall aim is to keep as much speech in the corpus as possible and to try to avoid the need for deleting recordings from the corpus due to some extra noises, dysfluencies, etc.
- 5. The conventions comprise both mandatory and optional transcriptions. All transcriptions should precisely follow the mandatory guidelines. The optional transcriptions are marked OPTIONAL in this document, and if provided should be documented and should follow these guidelines precisely. This is to regulate the task of external validation. Markings which

are optional have been chosen to be easily removed or translated by automatic means to yield the base transcription form.

6. The documentation provided with the database transcriptions should accurately provide details of which optional transcriptions were performed, and all relevant additional information, such as standard dictionary, preferred spelling variants, etc.

In summary, the principles are "Keep it simple" and "Document everything adequately".

M.2 Markings required for scoring

M.2.1 Case

Transcriptions are "case sensitive" in that case must be used to correctly encode information in the transcription, but the normal case rules of the language will not generally apply. It is recommended to use *lower case* only with no capitalisation of sentence initial text.

Note: This differs from the ATIS convention which did not distinguish the English indefinite article "a" from the spelled letter "a".

The only exceptions to this rule concern *proper nouns* and *spellings*. If the transcriber is certain that a word is a proper noun, the word should be *capitalised*.

Note: Per language decisions must be made concerning the treatment of compound proper nouns like 'Stephenson Way', 'mister Giscard d'Estaing', etc. These should be documented in each case.

M.2.2 Spelling

Normal lexical items will be represented by their spellings in the normal way. It is advised to stick to the normal spelling as much as possible. This also means that hyphens are used in the normal way. One dictionary or word list should be chosen (e.g. Duden for German, Larousse for French, Van Dale for Dutch). Each site/language maintains a lexicon of spellings of words used in the SPEECHDAT corpus. This file will be included on the CD-ROMs.

In many languages there are words or expressions which can be spelled in two or more different ways. To maintain consistency, each site/language must compile a list of such items, with the normalised spelling. For instance, in American English the spelling forms "all right" and "alright" coexist; one of these forms must be established as the standard.

It is probably profitable to always select the form yielding least 'words', because that should yield the most powerful language model. There is however a small technical advantage in having the norm as the multiple

word variant as spelling checkers can identify the single word forms very easily and convert them to the multiple word form automatically.

Abbreviations should be represented by their full orthographic forms, unless they are spoken in their abbreviated form. Exceptions are normally occurring abbreviations such as Mr, Mrs, Messrs, some of which do not have non-abbreviated forms.

To support homogeneity in spelling conventions used it is strongly recommended to employ an electronic spelling checker. If that is done, the make and type of the checker should be reported.

OPTIONAL:

An orthographic transcription means that the standard spelling in a given language is used for the symbolic representation of the speech. It is possible to include, a very restricted number of markings for regular variations in pronunciation. These cases must be clearly documented! Not more than two or three regular variations must be indicated.

For example: The absence of liaison in French may be indicated (see the final section on some language specific issues).

M.2.3 Number sequences

Number sequences (flight numbers, times, dates, aircraft types, money amounts, etc.) will be spelled out to reflect what was said (*flight six one three; seven thirty; august twenty first; seven forty seven; four hundred and ten dollars.*) If digits have alternate pronunciation forms (e.g. *zero* or *oh* or *naught* in English), the transcription should accurately reflect the form actually pronounced.

M.2.4 Letter sequences

Letter sequences occur in spelled words, ZIP-codes, acronyms and abbreviations ("D F W"; "A P slash eighty"; "P M"; "C O"; etc.) Letters should be in upper case, separated by a space. If letters have several names, like Y in Dutch, the actual name used must be transliterated instead of "Y". The AM and PM of times (e.g. "five thirty P M") will be treated as examples of letter sequences, i.e. upper case and separated by a space, with no periods. If a speaker pronounces letters, acronyms or abbreviations as a word, for example "British Rail" for BR, then these should be spelled out as words. If a speaker realises letters, especially consonants, by producing their phonetic form, upper case letters within slashes are used, e.g. /B/ /A/ /L/ spelling the Dutch word 'bal'.

As there can be different pronunciation schemes for letter spellings which it would be beneficial to identify, it is suggested that A B C...Z be used for the most common spelling form, and variations be marked by unique letter sequences which are not confusible with words, e.g. ZEE for US Z (cf. the

final section on language specific issues).

M.2.5 Punctuation

No punctuation will be provided in the transcription other than those symbols used for special transcription purposes. However the documentation should retain all punctuation provided to the speaker in the prompting text, including mistakes if these occurred.

M.2.6 Mispronunciations

Obviously mispronounced words that are nevertheless intelligible will be marked with one star before the word which is mispronounced, e.g. '*transportation' when 'transportetation' was said. Words preceded by a star include mispronunciations such as words with extra or omitted syllables, but a star should not be used to indicate pronunciations of words that represent normal dialectal (e.g. "warshed" for "washed" or "cah" for "car") or stylistic variation (e.g. "bout" for "about" or "wanna" for "want a" or for "want to". If the speaker would not consider the pronunciation an error, the star notation should not be used. Obviously, there may be some clear and some unclear cases; transcribers should use their best judgment. For languages with extremely long words (like German), it is possible to

put a star in the middle of these words if they are mispronounced. Note: This differs from the ATIS convention which used two stars.

M.2.7 Unintelligible words

Stretches of speech that are completely unintelligible are denoted by a sequence of two asterisks: "**".

Note: The corresponding ATIS marking was: [unintelligible]

M.2.8 Word fragments

Word fragments, i.e. instances in which the speaker did not complete a word, will be marked with a star following the fragment. As much of the word as is audible will be transcribed.

Ex.: please show fli* flights from Dallas

Fragments include cases in which only an initial consonant or vowel is heard:

Ex.: please show f* flights from Dallas

M.2.9 Verbal deletions

Verbal deletion means words spoken by the user but which, in the opinion of the transcriber, are superseded by subsequent speech explicitly (e.g.

"show flights I mean fares") or implicitly (e.g. "show me the fares flights to Boston"). Verbal deletions are not indicated as such in the SPEECHDAT corpora; transcribers should simply indicate all the words they hear. Verbal deletions occur any time there is a repetition or restart. In repetitions, one or more words are repeated, and there may or may not be extra material inserted into the repetition:

Ex.: show me the flights the flights to Boston

Ex.: show me the flights the nonstop flights to Boston

In restarts, words are not repeated, but the speaker changes direction, as in:

Ex.: show me the how many flights go to Boston

OPTIONAL:

As in the ATIS corpus, it is possible to indicate the verbal deletions by means of angle brackets. Each word in a verbal deletion should then be enclosed in angle brackets, for example:

Ex.: show me < the > < flights > the flights to Boston

Ex.: show me < the > < flights > the nonstop flights to Boston

Explicit phrases indicating the verbal deletion may be indicated as follows:

Ex.: show me <the> <flights> <<i>><<mean>> the nonstop flights to Boston

M.2.10 Non-speech acoustic events

Acoustic events enclosed in square brackets can come from the following set:

- FILLED_PAUSE Note that the filled pauses represent acoustic events similar acoustically and phonetically to speech. If used, transcriptions of hesitation sounds should be taken from an agreed list and put in the documentation file (e.g. [uh], [um], [er], [ah], [mm])
- SPEAKER_OTHER meaning all types of noises made by the speaker him or herself, like grunt, throat_clear, tongue_click, lip_smack, mouth_noise, loud_breath, laugh, cough, loud_sigh.

Note: Acoustic events such as inhalation, exhalation, tongue clicks, lip smacks, and breath noise will not be transcribed if they are low level and non-intrusive.

NONSPEAKER_OTHER meaning all types of noises not made by the speaker himself, like phone ringing, paper_rustle, door_slam, other_voices, TV-radio, ${\rm crosstalk}.$

Note: These will not be transcribed if they are low level and non-intrusive.

As these are currently in English partners are free to provide the language equivalents of these phrases. They should use only alphabetic characters and underscore, no spaces.

These events must be marked in the correct location in a transcribed utterance. It is often difficult to localise these events; transcribing the utterance first, and listening for these events in a second pass is the correct procedure. For noise events that occur over a span of one or more words, the transcriber should indicate the beginning of the noise, before the word it affects:

Ex.: "show the [nonspeaker_other] flights to Boston

Note: If a need is discovered to notate specific events these can be added, provided that they are defined clearly in the documentation and followed consistently. They should be capable of being remapped later.

Note: There is no notation for spans of noise events as in the ATIS conventions.

M.2.11 Prosodic annotation

Prosodic annotation is not obligatory for the SPEECHDAT corpora.

OPTIONAL:

Lengthening of sounds may be indicated by a colon:

Ex.: show me the f:lights to Boston

Silent pauses may be marked with a full stop surrounded by spaces ("."). This indicates a significant silence, i.e. one that is clearly noticeable by listening, and which is significantly longer than a silence associated with a stop consonant closure for the rate of speech used by the speaker.

Example: show me the . flights to Boston

M.2.12 Truncated waveforms

If a speech signal file is truncated due to a recording error, the following notation is to be used:

Beginning of utterance truncation:	\tilde{t} ranscription
End of utterance truncation:	$\operatorname{transcription}$
Beginning and end of utterance truncation:	$\tilde{transcription}$

There is a difference between an utterance which is truncated and is now incomplete, but which has not damaged the initial or final words, and an

utterance when word(s) have been damaged. The $\tilde{}$ indicates truncation of the word it is attached to. Otherwise truncated but good utterances will not be marked in any way.

Note: The tilde symbols "~" must not be separated from the truncated word by a white space (different from the ATIS convention).

M.3 Assessment (optional)

Following the Dutch POLYPHONE experience, each utterance may be assessed, independently of (and after completion of) the transcription. This information will be stored in a separate log file. Four different assessment types will be used:

1. GARBAGE

A file is to be marked as "Garbage" if

- it contains only background noise,
- it contains noises produced by a non-cooperative subject.

Garbage files should not be retained.

3. NOISE

Files are marked "Noise" if they contain clearly audible background noise in addition to the speech. A 'hard' criterion triggering this rating can be the failure of the recording platform to stop recording after the speaker completed the utterance (the platform can be set to consider two seconds of 'silence' as end of utterance).

4. OTHER

Files are marked "OTHER" if they contain

- one or more dysfluencies, hesitations or stuttering,
- an exceptional pronunciation of a word,
- speech that is (partly) unintelligible.
- 4. OK

Files were rated OK in all other conditions. Note that OK does not mean that the subject adhered exactly to the prompting text in read items; if he did not hesitate in speaking something else and there is no high level background noise, the item is rated OK. Also, utterances trimmed at the beginning or end are rated OK, provided that the first or last word present in the file are in no way damaged.

If a file can be rated both as "NOISE" and "OTHER", it must be rated "OTHER". Each file must be given exactly one rating.

It can be useful to provide an opportunity to make a comment about speaker characteristics (which could be stored once for all calls by that speaker) which are helpful for later analysis and selection of utterances, e.g. foreign or non-native speaker accent; very unclear, quiet or loud speakers; and especially stuttering or other significantly serious production characteristics, significantly poor voice quality, or uncooperative speakers whose data is not useful for training or testing. This could be optionally marked here and later included in the log file for the transcriptions or in the speaker description file.

M.4 Procedure

In Dutch POLYPHONE it proved advantageous to present the transcribers with the prompting text, whenever that was unambiguous. In those items the task of the transcriber was to decide whether the subject had adhered to the prompt without making additional noises. If the speaker deviated from the prompting text, or if there were extra noises, the default transcription had to be edited accordingly. If there were no discrepancies between prompt and speech, the default transcription could be added to the transcription file by hitting <CR>.

M.5 Speaker information

Speaker information should at least contain the following data:

- 1. Speaker/caller unique number,
- 2. sheet number,
- 3. sex,
- 4. age/age group/date of birth,
- 5. area/region (from a finite set, maximal 20 regions).

The following information is optional:

- Cordless phone,
- education level,
- size/weight of the speaker.

This information can appear once in the separate Speaker Description file which is itself referenced in each utterance label file. This avoids repeating this information in each label file. Some speaker information could be put in the label file, where we already have the row "SPI: sex, age, language". It is essential that the table provides a unique mapping of information for each caller. If the speaker and sheet numbers are the same, then they are simply provided as such, but both fields must be provided. If several speakers end up calling and using the same sheet number, then there must still be some code or number which identifies that speaker uniquely and relates the other information to them.

M.6 Some language-specific transcription options

M.6.1 French

The frequently occurring schwa at the end of words ending in a consonant is indicated. For the extra syllable at the end of the word curly brackets are used, e.g. neuf{e}.

The absence of liaison in French is indicated by adding a + sign behind the consonant. No marking is given if liaison has been applied, but if not the + sign is put after the consonant, e.g. "petit enfant" (pronounced with /t/: /pətit ãfã/), "petit+ enfant" (pronounced without /t/: /pəti ãfã/).

Letter sequences are presented like A B C D... Variations are indicated, e.g. W pronounced as V double (more regular is double V and this is simply indicated as W).

The [nonspeaker_other] events are indicated as laughter, coughing, throat_clear, telephone_ringing, paper_rustle, click, bip, door, chair, noise. These are put between square brackets.

M.6.2 English

Letter sequences in English. Known English spelling variants are:

- /dZeI/ vs. /dZaI/ (Scottish) for J
- /eItS/ vs. /heItS/ (aspirated) for H
- + /zed/ vs. /zee/ (USA) for Z

It is sufficient to use A B C ... Z for English and indicate less common pronunciations with special codes, e.g. ZEE.

M.6.3 Italian

For letter sequences, CSELT has designed a mapping between letters and graphemes used to transcribe letter spellings, see Table M.1.

For some letters there are more than one possibility, it depends on what the speaker said. Phonemes do not need solidus lines, because similar words do not exist.

Table M.1: CSELT mapping between letters and graphemes

LETTER	DEFAULT	ALTERNATE PRONUNCIATION
	SYMBOL	AND SYMBOL
А	А	-
В	BI	В
С	CI	C (CERA), CK (CANE)
D	DI	D
Е	Ε	E
F	\mathbf{EFFE}	F
G	GI	G (GELO), GH (GOLA)
Н	ACCA	_
Ι	Ι	-
J	I_LUNGO	I_LUNGA, GEI, IOTA
Κ	CAPPA	-
L	ELLE	L
Μ	EMME	М
Ν	ENNE	Ν
Ο	Ο	-
Р	PI	Р
Q	QU	-
R	ERRE	R
S	ESSE	S (SANO, CASA)
Т	ΤI	Т
U	U	-
V	VU	VI, V
W	VU_DOPPIO	VU_DOPPIA, DOPPIO_VU,
		DOPPIA_VUVI_DOPPIO,
		VI_DOPPIA, DOPPIO_VI, DOPPIA_VI
Х	ICS	-
Υ	IPSILON	IUPSILON, I_GRECO, I_GRECA
Z	ZETA	Z (RAZZA, ORZO)

N The Bavarian Archive for Speech Signals

N.1 Introduction

The Bavarian Archive for Speech Signals (BAS) was founded in January 1995 as an initiative of the Institute of Phonetics at the University of Munich, Germany.

The BAS will develop, validate, administrate and disseminate corpora of spoken German to the speech community as well as to the speech engineering industry.

N.2 General information

The BAS is located at the University of Munich, Germany.

BAS c/o Institut fuer Phonetik Schellingstr. 3/II 80799 Muenchen Germany Phone: +49-89-21802758 Fax: +49-89-2800362 email: bas@sun1.phonetik.uni-muenchen.de

Detailed and up-to-date information about the work of BAS and the available speech corpora can be found in the following URL:

http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html

N.3 Corpora

Presently the following German speech corpora are available on ISO 9660 CD-ROM:

- SIEMENS 1000 SI1000 (5 CD-ROMs) The corpus contains read speech of 10 different speakers. Each speaker has read approx. 1000 sentences from a German news paper corpus, thus resulting in a total of approx. 10000 recorded utterances.
- SIEMENS 100 SI100 (7 CD-ROMs) The corpus contains read speech of 101 different speakers. Each speaker has read approx. 100 sentences from either the SZ subcorpus or the CeBit subcorpus. The language is German. The subcorpus SZ contains 544 sentences from newspaper articles. The subcorpus CeBit contains 483 sentences from newspaper articles about the CeBit 1995. Each subcorpus is divided into 5 parts of approx. 100 utterances each.

Every speaker read only one part of one subcorpus (with some exceptions), thus resulting in a total of approx. 10100 recorded utterances.

- PHONDAT 1 PD1 (4 CD-ROMs, 2ND EDITION) The corpus contains read speech of 201 different speakers. Each speaker has read a subcorpus of 450 different sentences (including alphanumericals and two shorter passages of prose text); 8 speakers have read the whole sentence corpus. The speakers were recorded at four different sites in Germany (University of Kiel, University of Bonn, University of Bochum, University of Munich). The language is German. The corpus contains a total of 21681 recorded utterances.
 - Phonological segmentation by hand of small subset
 - Automatic alignment of the whole corpus
 - Documentation
 - Software
- PHONDAT 2 PD2 (1 CD-ROM, 2ND EDITION) The corpus contains read speech of 16 different speakers. Each speaker has read a corpus of 200 different sentences from a train inquiry task. The speakers were recorded at three different sites in Germany (University of Kiel, University of Bonn, University of Munich). The language is German. The corpus contains a total of 3200 recorded utterances.
 - Phonological segmentation by hand
 - Automatic alignment
 - Word segmentation
 - Prosodical segmentation
 - Documentation
 - Software

VERBMOBIL Spontaneous speech recorded in a dialogue task (appointment scheduling)

- VM CD 1.0.3 VM103 (1 CD-ROM, new edition in preparation)
 63 Dialogues 209 Appointm. 1840 Turns
- VM CD 2.0 VM20 (1 CD-ROM, new edition in preparation) 81 Dialogues 227 Appointm. 1538 Turns
- VM CD S 1.0 VMS10 (1 CD-ROM)

26 Free Dialogues (with overlap, stereo recordings) 2227 Turns

See the following URL for more information about the VERBMOBIL project: http://www.dfki.uni-sb.de/verbmobil/

STRANGE CORPUS 1 – SC1 (ACCENTS)(1 CD-ROM) The story "Nordwind und Sonne" read by 72 speakers with foreign accents and 16 native German speakers. The utterances read by the latter are phonologically segmented by hand.

N.4 Corpora in preparation

The following corpora will be available in the near future:

STRANGE CORPORA – SC Reference Corpora that reflect certain well known problems in speech processing, like accents, repair, breaks, hesitations, rep-
836 Part IV: Spoken language reference materials

etitions, extreme F0, background noise, pathological speech, speaker adaptation.

BAS EDITION OF VERBMOBIL CORPORA - VM 2nd extended edition

ARTICULATORY DATA - AD EMA data of speakers of SI1000 corpus

ERBA 10000 utterances from a train inquiry task

- SPINA ROBOT CONTROL TASK 10 sentences and 62 commands from the robot control domain spoken by 22 speakers in 5 versions, phonological segmentation, word segmentation (sentences).
- TED Transnational English Database

N.5 Other activities

BAS is currently developing tools for the automatic annotation and segmentation of very large speech corpora. This includes the automatic detection of variants of pronunciation, a statistical based alignment and a rule-based refinement of the outcome. This approach has two major goals:

- The first is to provide better and consistent annotations for speech science and engineering.
- The second is to get a statistical survey of phonetic word forms in German as a basis of an empirically based Complete Phonetic Theory (CPT).

More information about these activities can be found in the above mentioned URL.

N.6 Cooperations

The BAS seeks to cooperate with public institutions as well as with industrial partners to further develop new German speech databases. BAS can be a platform to redistribute existing German speech corpora or to unify forces for the development of new resources.

N.7 Future plans

Presently the BAS corpora are distributed on CD-ROM in a traditional way. For the second phase (starting 1997) we plan to change this policy into an on-line access to the BAS data using fast network technology.

N.8 More information

Please refer to the BAS WWW HomePage (see above) or contact us via email:

bas@sun1.phonetik.uni-muenchen.de

- ACCEPTANCE Decision outcome which consists in responding positively to a speaker (or speaker class) verification task.
- AGENT In the context of interactive systems, "agent" usually refers to a DIALOGUE PARTICIPANT, that is, the dialogue system or the user. However, it may also be used to refer to a human operator who takes over when a telephonebased dialogue goes wrong ("Please hold on; this call will be transferred to an agent").
- ALIGNMENT In determining the performance of a continuous speech recognition system, the response of the recogniser has to be compared to the transcription of the utterance presented to the system. In this process, the two word strings have to be *aligned* in order to compare them.
- ANALYTIC TESTING Procedure in which the listener is instructed to evaluate specific aspects of the performance of a speech output system, e.g. suitability of tempo, quality of segments, appropriateness of word stresses, sentence accents, etc.
- ANTONYMY Two words are antonyms (a) if they are co-hyponyms with respect to given meanings, and (b) if they differ in meaning in respect of those details of the same meaning which are not shared by their hyperonym.

Example: *manual* and *novel* are antonyms. Note that the term is sometimes restricted to binary oppositions, e.g. *dead – alive*.

- APPLICANT SPEAKER The speaker using a speaker recognition system at a given instant. Alternative terms: current speaker, test speaker, unknown speaker. This term can be ambiguous in certain contexts, as it may also be understood as a speaker who is unknown to the system. Though it is frequently found in the literature, we do not recommend to use it.
- APPLICATION DOMAIN An application domain is a particular DOMAIN which a dialogue system may be applied to/used in (for example, training for air-traffic controllers, timetable information provision, etc.)
- ASSESSMENT (of a recognition system) The process of determining the performance of the system, and evaluation of the use for a particular application.
- AUDIOTEX A system which plays pre-recorded messages to telephone callers provides an Audiotex service. The purpose of such services is to inform (e.g. weather forecasts, traffic information, etc.) or to entertain (e.g. horoscopes, joke lines, etc). Audiotex services are usually made available with Premium Rate Tariffs. Audiotex services tend to be tightly regulated, and they are not available in some countries.
- AUTOMATED SPEECH OUTPUT TESTING Speech output assessment procedure in which the human observer (listener in the case of audio output, or linguist in the case of symbolic output) has been replaced (modelled) by an algorithm. Automated assessment presupposes that we know exactly how human observers evaluate differences between two (acoustic or symbolic) realisations of the same linguistic message.

- AUTOMATIC SPEECH RECOGNITION SYSTEM A device that can recognise the human's speech, and can output the words that are spoken by a human.
- BASELINE REFERENCE (CONDITION) Speech output of a system that contains no specific intelligence.
- BENCHMARK The value that characterises some reference system against which a newly developed system is (implicitly) set off.
- BENCHMARK TEST An efficient, easily administered test, or set of tests, that can be used to express the performance of a speech output system (or some module thereof) in numerical terms.
- BLACK BOX APPROACH Performance evaluation of a system as a whole, typically used to compare systems developed by different manufacturers, or to establish the improvement of one system relative to an earlier edition (comparative testing). Black box evaluations consider the overall performance of a system without reference to any internal components or behaviours. Evaluations of this kind address large questions such as "How good is it as an *integrated system*?" rather than detailed questions of the "What is its word recognition rate?" variety. Compare GLASS BOX APPROACH.
- CANNED SPEECH Speech which has been recorded for use in the prompts or information play-outs of a dialogue system is referred to as canned speech or canned messages. A number of canned messages can be played out one after the other to create a single system utterance. For example, the following system utterance consists of X canned messages (identified by <...>) concatenated together: "<Flight> <XY> <five> <seven> <two> <from> <London> <to> <Brussels> <will arrive at> <fifteen> <thrufty> <seven>." With careful attention to prosodic issues canned speech can provide a high quality, natural-sounding interface. SpeecH synthesis, though less natural-sounding, is more flexible and thus more appropriate when lengthy or lexically rich system utterances are required.
- CATEGORICAL ESTIMATION Rating method where the subject has to assign to (some aspect of) a speech output system a value from a limited range of prespecified values, e.g. "1" representing extremely poor and "10" excellent intelligibility.
- CO-HYPONYMY Two words are co-hyponyms if and only if there is a word which is a hyperonym of each (in the same meaning of this word).

Example: manual and novel are co-hyponyms in relation to book.

- COMMON-PASSWORD SPEAKER RECOGNITION SYSTEM A text-dependent speaker recognition system for which all registered speakers have the same voice password.
- COMMUNICATION MEDIA Media (or "means") refer to materials or devices which are used by an interactive dialogue system to communicate with the user.
- COMMUNICATION MODES A mode refers to perception senses which allow for communication: the following modes may be identified: vocal, visual, auditive, tactile, olfactive.

COMPARATIVE TESTING See BLACK BOX APPROACH.

COMPETENCE (vs. performance) A technical term in theoretical linguistics. Competence is a speaker/hearer's knowledge of his own language. This is contrasted with PERFORMANCE, what speakers actually say. Thus, though my competence tells me that the past tense of the verb "go" is *went*, a host of factors including fatigue, distraction, or word-play may result in my performance production of the ill-formed **goed*.

- COMPREHENSION TEST Procedure testing a listener's understanding of a speech stimulus at the sentence or text level (often by asking the listener to answer content questions).
- CONCEPT-TO-SPEECH SYSTEM Speech output system that converts some abstract representation of a communicative intention to speech.
- CONTINUOUS SPEECH/CONNECTED WORDS A speaking style where the words form a continuous signal, i.e. the words follow each other fluently. Contrary to isolated words.
 - The distinction between "connected words" and "continuous speech" is somewhat technical. A connected word recogniser uses words as recognition units, which can be trained in a isolated word mode. Continuous speech is generally associated with large vocabulary recognisers that use *phones* as recognition units and can be trained with continuous speech.

CORRECTION RATE (CR) Percentage of all turns which are correction turns.

- CROSS-VALIDATION Cross-validation is a technique in statistical estimation by which the parameters of a model are optimised on a new unseen test set. In the context of stochastic language modelling, cross-validation is used to estimate the smoothing parameters.
- CUT-THROUGH The system hears and understands simultaneously (single step).
- DELETION (or miss) A word in the utterance that is not recognised.
- diagnostic testing See glass box testing.
- DIALOGUE ACT This term is favoured by some authors who wish to appeal to the basic idea of SPEECH ACTS without buying into the whole philosophical apparatus of Speech Act Theory. The basic idea is that utterances can be categorised into broad classes such as questions, confirmations, statements, etc. In keeping with the practical engineering view which usually informs analyses which use the dialogue act notion, the exact inventory of categories tends to be determined by the particular needs of each dialogue application.
- DIALOGUE GRAMMAR A grammar for describing a set of well-formed dialogues. The terminal symbols in a dialogue grammar are SPEECH ACT OF DIALOGUE ACT labels (though for convenience these labels may also be treated as the start symbol for more conventional sentence or utterance grammars). A dialogue grammar might, for example, contain a rule which says that a simple information request consists of two turns, the first of which is a question, and the second of which is an answer. The philosophical roots of dialogue grammars lie in the field of DISCOURSE ANALYSIS.
- DIALOGUE HISTORY A system-internal record of what has happened in a dialogue so far. The dialogue history provides the immediate context within which interpretation takes place.
- DIALOGUE MANAGER The component in an INTERACTIVE DIALOGUE SYSTEM which is responsible for maintaining dialogue coherence. Functions typically undertaken by a dialogue manager include the following:

- maintaining a model of the current dialogue context;
- interpreting utterances in context;
- linking interpretations to actions;
- thinking of something to say next;
- generating topdown predictions of the next user utterance;
- keeping track of who knows what;
- generating utterances which are cooperative;
- selecting an appropriate dialogue strategy;
- recovering from dialogue breakdowns.

DIALOGUE PARTICIPANT Each of the participants involved in a dialogue – those speaking and those listening – is a dialogue participant.

- DICTIONARY (or lexicon) A lookup-table of pronunciations of all the words a (continuous) speech recognition system is capable to recognise.
- DISCOUNTING Discounting is a technique in the context of language model smoothing by which the relative frequencies are *discounted* to allow for unseen events.
- DISCOURSE ANALYSIS The branch of linguistics which is concerned with the analysis of naturally occurring connected spoken or written discourse.
- DOMAIN The area of language usage for which a recognition system is designed to be used, a (possibly ill-defined) subset of general activity (such as business, avionics, aeronautics, medicine, transport, etc.) in which some coherent collection of TASKS may be carried out.
- ENVIRONMENT The environment is the total context in which a recognition or interactive dialogue system is located. For example, a dashboard control system operates in an in-car environment. Environments may be characterised in many different ways. Most commonly, however, factors which might affect the performance of the system (such as high background noise) are singled out to describe environments.
- ERROR RATE The fraction of errors made by a recognition system, i.e. the number of errors divided by the number of words to be recognised. Often expressed as a percentage. See also DELETION, SUBSTITUTION and INSERTION.
- ESTIMATOR A mathematical expression that can be used to estimate the value of a statistical property, such as the mean or variance.
- EVENT-DEPENDENT SPEAKER RECOGNITION SYSTEM A text-independent speaker recognition system for which test utterances must contain a certain linguistic event (or class of events) while the rest of the acoustic material is discarded. This approach requires a preliminary step for spotting and localising the relevant events.
- EXCHANGE A pair of contiguous and related turns, one spoken by each party in the dialogue.
- FALSE (SPEAKER) ACCEPTANCE (Sometimes called type-II error) Erroneous acceptance of an impostor in open-set speaker identification or in speaker verification.
- FALSE (SPEAKER) REJECTION (Sometimes called type-I error) Erroneous rejection of a registered speaker or of a genuine speaker in open-set speaker identification or speaker verification.

- FIELD TESTING Speech output test procedure entirely run in the actual application, using the real-life situation with the actual end-users.
- FIXED-VOCABULARY SPEAKER RECOGNITION SYSTEM A text-independent speaker recognition system for which test utterances are composed of words, the order of which varies across speakers and sessions, but for which all the words are pronounced at least once by the speaker when he registers to the system.
- FLAWLESS SPEECH The unweighted reproducible 1:1 transduction of an acoustical signal emitted by a speaker into a sequence of 2 byte numbers that is free of any room or environment information, exhibits a sufficient signal-to-noise ratio of at least 50 dB, and has been produced under recording conditions that do not impose any stress upon the speaker in addition to what might be intended for a given talking situation.
- FORMAL LANGUAGE An invented language, usually developed for purposes of representation and manipulation (for example, in mathematics, logic or semantics) and not for purposes of communication.
- FUNCTIONAL TESTING Assessment of speech output in terms of how well a system actually performs (some aspect of) its communicative purpose.
- GENUINE SPEAKER A speaker whose real identity is in accordance with the claimed identity. By extension: a speaker whose actual character and claimed class are in accordance. (For instance, a female speaker claiming that she is a female speaker, in sex verification.). Alternative terms: authentic speaker, true speaker, correct speaker.
- GLASS BOX APPROACH Test methodology in which the effects of all modules in a text-to-speech system but one are kept constant, and the characteristics of the free module are systematically varied, so that any difference in the assessment of the system's output must be caused by the variations in the target module (diagnostic testing). Glass box testing presupposes that the researcher has control over the input and output of each individual module. Compare BLACK BOX APPROACH.
- GLOBAL TESTING Procedure in which the listener is instructed to attend to the general performance of a speech output system, e.g. in terms of listening effort, acceptability, and naturalness.
- GRAMMAR A set of rules that define how the words in a language can follow each other. This can include information about the probability that a sequence of words occurs.
- GRAPHEME-PHONEME CONVERSION Module within a text-to-speech system that accepts a full-blown orthographic input (i.e. the output of a preprocessor), and outputs a string of phonemes (often, but not necessarily) including (word) stress marks, (sentence) accent positions, and boundaries.
- HETEROGRAPHY Two orthographic forms of the same word are heterographs. Example: *standardise* - *standardize* /st{nd@daIz/.
- HETEROPHONY Two phonological forms of the same word are heterophones. Example: either /aID@/ - /i:D@/ 'disjunction'.
- HOMOGRAPHY Two words with the same orthographic form and different phonological forms are (heterophonic) homographs.

Example: row /roU/ 'horizontal sequence', /raU/ 'noise, quarrel'.

HOMONYMY Two words with the same orthographic and phonological forms, but different syntactic categories and/or meanings are homonyms.

Example: mate /meIt/ 'friend' or 'state of play in a chess game'.

HOMOPHONY Two words with the same phonological form and different orthographic forms are (heterographic) homophones.

Example: meet /mi:t/ 'encounter' - meat /mi:t/ 'edible animal tissue'.

- HUMAN-COMPUTER INTERACTION Often abbreviated to HCI. Any interaction between a person and a computer. Some writers use *human-computer dialogue* as a synonym for HCI, while others use it to identify a subtype of HCI in which natural language is used as the primary or the only medium of communication. A genuine synonym for HCI is man-machine interaction (MMI).
- HUMAN-HUMAN INTERACTION Any encounter between two (or more) people is a human-human interaction. Thus, a conversation is a human-human interaction. Human-human interactions are interesting to interactive dialogue technologists because of the light they may shed on HUMAN-COMPUTER IN-TERACTIONS. However, a body of findings is being growing which shows that human-human and human-computer natural language dialogues differ systematically. Lessons for system design based on human-human dialogues must be interpreted in the light of these.
- HYPERONYMY If the meaning of one word is entailed by the meaning of another, it is a hyperonym of the other (a superordinate term relative to the other). Example: *book* is a hyperonym of *manual* as the meaning of *book* is implied by the meaning of *manual* (in one of its meanings).
- HYPONYMY The converse of hyperonym. If the meaning of one word entails the meaning of another, it is a hyponym of the other (a subordinate term relative to the other).

Example: *manual* is a hyponym of *book* as the meaning of *manual* implies the meaning of *book*.

- IDENTIFICATION TEST Procedure by which the listener is asked to identify a speech stimulus in terms of some (closed or open) set of response alternatives (e.g. some or all of the phonemes in the language).
- IDENTITY ASSIGNMENT Decision outcome which consists in attributing an identity to an applicant speaker, in the context of speaker identification. For speaker classification, the term *class assignment* should be used instead.
- IMPOSTOR In the context of speaker identification, an impostor is an applicant speaker who does not belong to the set of registered speakers. In the context of speaker verification, an impostor is a speaker whose real identity is different from his claimed identity. Alternative terms: impersonator, usurpator. (Both terms are very rarely used.) For speaker classification tasks, this concept is better rendered by the term *discordant speaker* (for instance, a child claiming that he is an adult, in age verification).
- INSERTION (or false alarm) The response of a word that was not in the utterance presented to a speech recognition system.

INTERACTION Communication of information between two AGENTS, in which (ex-

cept for the special case of the initial TURN) an agent's contribution at any given point can be construed as a response to the previous turn or turns.

- INTERACTIVE DIALOGUE SYSTEM A computer system capable of engaging in turnby-turn communication with a human user. In the general case, communication between the person and the system could use any COMMUNICATION MODE or MEDIUM (or several simultaneously). In this chapter, however, the term is usually used more restrictively to apply to systems whose primary mode of communication is spoken natural language. See also INTERACTIVE VOICE RESPONSE and SPOKEN LANGUAGE DIALOGUE SYSTEM.
- INTERACTIVE VOICE RESPONSE (IVR) Interactive Voice Response (IVR) is what the commercial world calls interactive dialogue. As such, its scope encompasses certain kinds of simple interaction which research scientists do not normally think of as dialogues. For example, a telephone caller calling a weather forecasting AUDIOTEX service might be asked to say one of the words "today", "tomorrow" or "weekend". In the basis of what the system recognises, a *canned* weather forecast will be played. This is an example of IVR which is also widely known as Voice Response (VR), and a system which supports VR is usually known as a Voice Response Unit (VRU).
- INTERPOLATION Interpolation or linear interpolation is a technique in the context of language model smoothing by which the relative frequencies of a specific model are *interpolated* with those of a more general model. The term interpolation is often synonymous with smoothing.
- ISOLATED WORDS A speaking style where the words (or small phrases) are uttered separately, with small pauses in between. Contrary to continuous speech.
- JUDGMENT TESTING Procedure whereby a group of listeners is asked to judge the performance of a speech output system along a number of rating scales. (also called opinion testing in telecommunication research)
- LABORATORY TESTING Speech output test procedure entirely run in a laboratory, either abstracting from in vivo complications or trying to simulate real-life situations.
- LANGUAGE MODEL A language model in speech recognition is used to improve the recognition accuracy. Its task is to capture the redundancy inherent to the word sequences to be recognised. This redundancy may result from both the task specific constraints and general linguistic constraints.
- LEAVING-ONE-OUT Leaving-one-out is a special kind of cross-validation where no additional test set is needed. Instead it is generated from the training observations by leaving out one observation at a time.
- LINGUISTIC INTERFACE First part of a text-to-speech system, which transforms spelling into an abstract phonological code (which in turn is converted to sound by the acoustic interface). The linguistic interface includes text preprocessing, grapheme-phoneme conversion, assignment of (word) stress, (sentence) accent, and boundary positions, and choice of intonation pattern.
- LOMBARD-EFFECT The effect that humans speak at a higher level (use more vocal effort) in conditions of higher environmental noise.
- m-GRAM MODEL An m-gram model is a stochastic language model that is based on conditional probabilities depending only on the (m-1) immediate prede-

cessor words.

- MAGNITUDE ESTIMATION Rating method where the subject is presented with an (auditory) stimulus and is asked to express the perceived strength/quality of the relevant attribute (e.g. intelligibility) numerically ("type in a value") or graphically ("draw a line on the computer screen").
- (SPEAKER) MISCLASSIFICATION Erroneous identity assignment to a registered speaker in speaker identification.
- MISTAKEN SPEAKER The registered speaker owning the identity assigned erroneously to another registered speaker by a speaker identification system.
- MODALITIES Modalities concern the way a communicating agent/party uses a mode: for speech, different modalities may be identified, whether continuous speech or isolated words are used, whether a whispering or shouting style is used, etc.
- MORPHOLOGICAL DECOMPOSITION Analysis of orthographic words into morphemes, i.e. elements belonging to the finite set of smallest subword parts with an identifiable meaning. Morphological decomposition is necessary when the language/spelling allows words to be strung together without intervening spaces or hyphens.
- NATURAL LANGUAGE Any non-invented language is a natural language. Thus, even the language used between people and invented systems can be termed "natural" if it is what users spontaneously produce in response to the situation. Natural languages can be contrasted with FORMAL LANGUAGES. See also RESTRICTED LANGUAGE and SUBLANGUAGE.
- OFF-LINE TESTING Procedure in which subjects are given some time to reflect before responding to a (spoken) stimulus.
- ON-LINE TESTING Procedure that requires an immediate response from the subjects, tapping the perception process before it is finished.
- OPINION TESTING See JUDGMENT TESTING.
- ORAL DIALOGUE See SPOKEN LANGUAGE DIALOGUE. This term is quite widely used, though it is less favoured by native speakers of English than by those who have learned it as a second-language.
- PAIRED COMPARISON A psychophysical method. It is used when subjects are required to judge between two stimuli. In LES this might be judging which of two recogniser outputs has more or less intelligibility.
- PARAMETRIC AND NON-PARAMETRIC TESTS A distinction between two basic forms of statistical tests employed in simple hypothesis testing. Parametric tests are used when continuous measures are available. Non-parametric tests are used otherwise.

PARTY See DIALOGUE PARTICIPANT.

PERFORMANCE (vs. competence) A term from theoretical linguistics to describe what speakers actually say. This is contrasted with COMPETENCE, what speaker/hearers know about their language. It is generally held that there is a dislocation between competence and performance such that there is not a straightforward mapping from one to the other.

PERFORMANCE EVALUATION See BLACK BOX APPROACH.

PERPLEXITY A measure for the complexity of a grammar.

- PERPLEXITY The (corpus) perplexity is a quantitative measure of the redundancy (or difficulty) of a recognition task for a given text corpus and a given language model. It measures how well the word sequences can be predicted by the language model.
- PERSONAL-PASSWORD SPEAKER RECOGNITION SYSTEM A text-dependent speaker recognition system for which each registered speaker has his own voice password.

PHONE A subword unit of speech that represents a particular sound.

- PHONETICALLY BALANCED SENTENCES Sentences containing phonemes according to their frequency of occurrence in a given language.
- PHONETICALLY RICH SENTENCES Sentences containing approximately uniform phoneme frequency distributions.

POPULATION The collection of all objects that are of interest for the task in hand.

- PROSODY Those properties of speech utterances that cannot be derived in a straightforward fashion from the identity of the vowel and consonant phonemes that are strung together in the linguistic representation underlying the speech utterance, e.g. intonation (i.e. speech melody), word and phrase boundaries, (word) stress, (sentence) accent, tempo, and changes in speaking rate.
- REGISTER A term from sociolinguistics which is used to identify a language variety according to its use. Every speaker of a NATURAL LANGUAGE has command of a multitude of different registers. For example, the variety of language used in a social gathering with old friends is very different to that used with a doctor in a medical surgery. Context of use can affect all aspects of language use: the choice of words, the kind of syntactic constructions, accent, etc. Register may vary during the course of a single interaction. So, for example, a very formal register may be used when people first meet but, as the conversation develops, a more relaxed and informal register may take over. For this reason it is inappropriate to try to identify what might be called "the register for human-computer dialogue" because such a thing is unlikely to exist as a unitary phenomenon. Instead, it is usual to try to model language over the range of varieties which might be used in some given application domain. This model is usually called a SUBLANGUAGE.
- REGISTERED SPEAKER A speaker who belongs to the list of registered users for a given speaker recognition system (usually a speaker who is entitled to use the facilities, the access of which is restricted by the system). For speaker classification systems, we propose the term *conform speaker* to qualify a speaker who belongs to one of the classes of speakers for a given speaker classification system. For instance, for a spoken language identification system that discriminates between languages spoken in Switzerland, a *conform speaker* is a speaker who speaks either German, French, Italian or Romansch, but not a language that the system does not expect. Alternative terms: reference speaker, valid speaker, authorised speaker, subscriber, client.
- REJECTION Decision outcome which consists in refusing to assign a registered identity (or class) in the context of open-set speaker identification or classifica-

tion, or which consists in responding negatively to a speaker (class) verification trial.

- RESTRICTED LANGUAGE A variety of NATURAL LANGUAGE which is restricted by externally imposed rules of use. These rules typically limit the vocabulary and the range of acceptable syntactic constructions. Restricted languages tend to be used in contexts where rapid, effective communication of a small set of basic facts is paramount, for example, in air traffic control. Because of the tightly constrained nature of restricted languages, they are seen by many to be good candidates for modelling in interactive dialogue systems. However, this advantage must be weighed against the safety-critical function of many such languages in real use.
- SAMPLE Typically, a measure cannot be taken on all units of a population. In these cases, a sample is taken. Provided precautions are taken as set out in the text, this sample may be used to study the variable of concern in the population.
- SEGMENTS Consonants and vowels of a language.
- SIGNAL DETECTION THEORY A model that may be used for studying the performance of speech recogniser performance. The basic idea behind signal detection theory is that errors convey information concerning how the system is operating (in this respect, it is an advance on simple error measures).
- SIGNAL-TO-NOISE RATIO The ratio of information-carrying signals (speech) to background noise. Expressed in dB.
- SMOOTHING Smoothing is a method that is needed in the context of stochastic language modelling to counteract the effect of sparse training data. The goal of smoothing is to guarantee that all probabilities are different from zero.
- SPEAKER CLASSIFICATION Any decision-making process that uses some features of the speech signal to determine some characteristics of the speaker of a given utterance.
- SPEAKER RECOGNITION Any decision-making process that uses some features of the speech signal to determine some information on the identity of the speaker of a given utterance.
- SPEAKER CLASS IDENTIFICATION Any decision-making process that uses some features of the speech signal to determine the class to which the speaker of a given utterance belongs.
- SPEAKER CLASS VERIFICATION Any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance belongs to a given class.
- SPEAKER IDENTIFICATION Any decision-making process that uses some features of the speech signal to determine who the speaker of a given utterance is.
- SPEAKER VERIFICATION Any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance is a particular person, whose identity is specified.
- SPEECH ACT A speech act is the informational action that a speaker effects by producing an utterance. For example, asking a question, offering information, and making a promise are three different types of speech act. The basic idea

of speech acts is vitally important in work on dialogue systems. Speech acts serve as the base level of categorisation for dialogue work (in much the way that word classes have that function at the lexical level). So, for example, DIALOGUE GRAMMARS can be written which describe well-formed sequences of speech acts.

Many researchers working on interactive dialogue systems wish to use the notion of speech act without enlisting the whole philosophical apparatus of Speech Act Theory (cf. Austin 1962; Searle 1969); for this purpose the term DIALOGUE ACT has been coined and is steadily growing in acceptability.

SPEECH OUTPUT ASSESSMENT See SPEECH OUTPUT TESTING.

- SPEECH OUTPUT SYSTEM Some artifact, either a dedicated machine or a computer programme, that produces signals that are intended to be functionally equivalent to speech produced by humans. In the present state of affairs speech output systems generally produce audio signals only, but laboratory systems are being developed that supplement the audio signal with the visual image of the (artificial) talker's face.
- $\ensuremath{\texttt{SPEECH}}$ output testing Determination of the quality of (some aspect(s) of) a speech output system.

 $\ensuremath{\mathsf{SPEECH}}$ output evaluation See speech output testing.

- SPEECH SYNTHESIS Speech Synthesis is the name given to the production of speech sounds by a machine. Most speech synthesisers take a text string as input and produce a spoken version of the text as output. Some systems allow the text string to be annotated with prosodic markers which result in changes to the intonational pattern of the speech produced.
- SPOKEN LANGUAGE CORPUS Any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use.
- SPOKEN LANGUAGE DIALOGUE Also known as ORAL DIALOGUE. A complete spoken verbal interaction between two parties (in the present case, a system and a human being), each of whom is capable of independent actions. A dialogue is composed of a sequence of steps which are, in some way, related and build on each other. Dialogue systems are thus more sophisticated than question/answer systems, in which one agent may pose a succession of unrelated queries to the other agent.
- SPOKEN LANGUAGE DIALOGUE SYSTEM A variety of INTERACTIVE DIALOGUE SYS-TEM in which the primary mode of communication is spoken natural language. Spoken language dialogue systems take human-human conversation as their inspiration, though differences are bound to persist into the forseeable future by virtue of the character of such systems as constrained designed artifact. Spoken language dialogue systems support a much more natural kind of dialogue than INTERACTIVE VOICE RESPONSE systems.
- SPOKEN LANGUAGE IDENTIFICATION Any decision-making process that uses some features of the speech signal to determine what language is spoken in a given utterance.
- SPOKEN LANGUAGE VERIFICATION Any decision-making process that uses some features of the speech signal to determine whether the language spoken

in a given utterance is a particular language.

- STOCHASTIC GRAMMAR A stochastic grammar is a stochastic language model that is based on a (context free) grammar; the grammar rules are assigned probabilities such that each word string generated by the grammar has a non-zero probability.
- STOCHASTIC LANGUAGE MODEL A stochastic language model is a language model that assigns probabilities to the allowed word sequences; typically all word sequences have a non-zero probability.
- SUBLANGUAGE The subpart of some NATURAL LANGUAGE which is deemed to be relevant to some given task and/or application domain. Interactive dialogue systems are not currently capable of modelling an average speaker's entire linguistic competence, so the normal approach is to identify and model only the sublanguage which is relevant to the function or functions which the interactive dialogue system is intended to perform. The idea of sublanguage is related to, but distinct from the linguistic notion of REGISTER. A sublanguage in the context of interactive dialogue systems should not be confused with a sublanguage in the mathematical sense. In the latter case, the language of which the sublanguage is a part is formally well-defined; in the former case it is not.
- SUBSTITUTION (or misclassification) A response of a recognised that is different from the word in the utterance presented to a recognition system.
- SYNONYMY Two words are synonyms if and only if they have the same meaning (or at least have one meaning in common), i.e. if the meaning of each entails the meaning of the other. They are partial synonyms if either has additional readings not shared by the other. They are full synonyms if they have no readings which is not shared by the other.

Example: *manual* and *handbook* are partial synonyms (*manual* is also, among other things, a term for a traditional organ keyboard). Full synonyms are rare. By implication, synonyms are also co-hyponyms.

- SYSTEM CORRECTION RATE (SCR) Percentage of all system turns which are correction turns.
- SYSTEM-IN-THE-LOOP A speech data collection method which involves getting subjects to use an existing spoken language dialogue system, and recording what they say.
- TASK A task consists of all the activities which a user must develop in order to attain a fixed objective in some DOMAIN.
- TASK-ORIENTED DIALOGUE A dialogue concerning a specific subject, aiming at an explicit goal (such as resolving a problem or obtaining specific information). For example, dialogues concerned with obtaining travel information or booking theatre tickets are task-oriented.
- TEXT-DEPENDENT SPEAKER RECOGNITION SYSTEM A speaker recognition system for which the training and test speech utterances are composed of exactly the same linguistic material, in the same order (typically, a password).
- TEXT-INDEPENDENT SPEAKER RECOGNITION SYSTEM A speaker recognition system for which the linguistic content of test speech utterances varies across trials.

- TEXT PREPROCESSING The first stage of the linguistic interface of a text-to-speech system, which handles punctuation marks and other non-alphabetic textual symbols (e.g. parentheses), and expands abbreviations, acronyms, numbers, special symbols, etc. to full-blown orthographic strings.
- TEXT-PROMPTED SPEAKER RECOGNITION SYSTEM A speaker recognition system for which, during the test phase, a written text is prompted (through an appropriate device) to the user, who has to read it aloud.
- TEXT-TO-SPEECH SYSTEM Speech output system that converts orthographic text (generally stored in a computer memory as ASCII codes) into speech.
- TOPLINE REFERENCE (CONDITION) Speech output that represents optimum performance, typically by a professional human talker.
- TRAINING The process in which a speech recognition system learns the pronunciation of words to be recognised at a later instance.
- TRANSACTION The part of a dialogue devoted to a single high-level task (for example, making a travel booking or checking a bank account balance). A transaction may be coextensive with a dialogue, or a dialogue may consist of more than one transaction.
- TURN A stretch of speech, spoken by one PARTY in a dialogue. A stretch of speech may contain several linguistic acts or actions. A dialogue consists of a sequence of turns produced alternately by each party. Turns are also known as utterances.
- UNPROMPTED SPEAKER RECOGNITION A speaker recognition system using totally spontaneous speech, i.e. for which the user is totally free to utter what he wants. (Here, a further distinction could be made between *language dependent* and *language independent* systems.), or for which the system has no control over the speaker. (For instance, in forensic applications, the speaker may not be physically present, or may not be willing to cooperate.).
- UNRESTRICTED TEXT-INDEPENDENT SPEAKER RECOGNITION SYSTEM A textindependent speaker recognition system for which no constraints apply regarding the linguistic content of the test speech material.
- USER CORRECTION RATE (UCR) Percentage of all user turns which are correction turns.
- VIOLATED SPEAKER The registered speaker owning the identity assigned erroneously to an impostor in open-set speaker identification. The registered speaker owning the identity claimed by a successful impostor in speaker verification.
- VOCABULARY The set of words that an automatic speech recognition system is capable to recognise.
- VOICE CHARACTERISTICS Those aspects of speech which remain relatively constant over longer stretches of speech, and constitute the background against which segmental and prosodic variation is produced and perceived (e.g. mean pitch level, mean loudness, mean tempo, harshness, creak, whisper, tongue body orientation, dialect).
- VOICE-PROMPTED SPEAKER RECOGNITION SYSTEM A speaker recognition system for which, during the test phase, the user has to repeat a speech utterance, which he listens to through an audio device.

VOICE QUALITY See VOICE CHARACTERISTICS.

 $\ensuremath{\mathsf{VOICE}}$ stop In a first step the system hears; it needs a second step to understand.

- WIZARD-OF-OZ SIMULATION Simulation of the behaviour of an interactive automaton by a human being. This can be done (i) by speaking to the user in a disguised voice, (ii) by choosing and triggering system predefined responses, (iii) by manually modifying some parameters of the simulation system, or (iv) by using a person to simulate the integration of existing system components (a bionic Wizard-of-Oz simulation).
- WORD GRAPH OR WORD LATTICE A word graph or lattice is used in the context of search in speech recognition to provide an explicit interface between the acoustic recognition and the application of the language model. The word graph or lattice should contain the most likely word hypotheses where in addition to the word hypothesis the start and end times, the nodes and an acoustic probability are given.

AD	Articulatory Data
ADPCM	Adaptive Differential PCM
AGC	Automatic Gain Control
AI	Articulation Index
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ANSI	American National Standards Institute
API	Application Programming Interface
ARPA	Advanced Research Program Agency
ARPA	Advanced Research Projects Agency
Ars	Adverse Recognition System
ASCII	American Standard Code for
	Information Interchange
ASG	Automatic Speech Generation
ASR	Automatic Speech Recognition
Atis	Air Line Travel Information System
ATM	Automatic Teller Machine
BAS	Bavarian Archive for Speech Signals
BNC	British National Corpus
BNF	Backus Naur Form (Backus Normal Form)
CA	Contextual Appropriateness
CART	Classification And Regression Trees
CD-ROM	Compact Disk–Read Only Memory
CFA	Confirmatory Factor Analysis
CLID	CLuster IDentification
COCOSDA	International Committee for Collaboration
	in Speech Assessment and Databases
CPT	Complete Phonetic Theory
CPU	Central Processing Unit
CR	Correction Rate
Cril	Computer Representation of
	Individual Languages
CTI	Computer Telephony Integration
CTS	Concept-To-Speech
CVC	Consonant Vowel Consonant
CVSD	Continuously Variable Slope Delay
DAT	Digital Audio Tape
DATR	Default ATtribute Representation
DBMS	DataBase Management System
DD	Dialogue Duration
DDL	Dialogue Description Language
DP	Dynamic Programming

DPSIE	Diagnostic Pairs Sentence Intelligibility
	Evaluation Test
DRA	Defence Research Agency
DRT	Diagnostic Rhyme Test
DSP	Digital Signal Processing
DTMF	Dual Tone Multi-Frequency
DTW	Dynamic Time Warping
DVI	Direct Voice Input
DVO	Direct Voice Output
EAGLES	Expert Advisory Group on
	Language Engineering Standards
EER	Equal Error Rate
EIN	Equivalent Input Noise
ELRA	European Language Resources Association
ELSA	ESPRIT Labelling System Assessment software
EM	Expectation-Maximisation
EMA	Electromagnetic Articulography
EMG	electromyograph
EUROPEC	European Program d'Enregistrement de Corpus
FRANCIL	Réseau Francophone de l'Ingénierie
	de la Langue
GLIM	Generalised LInear Modelling Techniques
GNE	Glottal-to-Noise Excitation
GUI	Graphical User Interface
HCI	Human–Computer Interaction
HENR	Human Equivalent Noise Ratio
HMM	Hidden Markov Model
HP	high pass filter
HPSG	Head-Driven Phrase Structure Grammar
IC	Integrated Circuit
ICASSP	International Conference on Acoustics, Speech
	and Signal Processing
ICPhS	International Congress of Phonetic Sciences
ICSLP	International Conference on Spoken Language
	Processing
ILEX	Integrated LEXicon
IPA	International Phonetic Alphabet
IPO	Instituut voor Perceptie Onderzoek
IVR	Interactive Voice Response
KRL	Knowledge Representation Language
LDC	Linguistic Data Consortium
LES	Language Engineering Standards
LKRL	Lexical Knowledge Representation Language

LOO	Leaving-One-Out
LP	low pass filter
LRE	Linguistic Research and Engineering Programme
LTI	Linear and Time-Invariant system
Lin-PCM	Linear Pulse Code Modulation
Log-PCM	Logarithmic Pulse Amplitude Modulation
MIME	Multipurpose Internet Mail Extensions
MMI	Man–Machine Interaction
MNRU	Modulated Noise Reference Unit
Morpa	MORphological PArser
MPI Test	Minimal Pairs Intelligibility Test
MRPA	Machine Readable Phonetic Alphabet
MRT	Modified Rhyme Test
MVIP	Multi-Vendor Integration Protocol
NERC	Network of European Reference Corpora
NIST	National Institute of Standards
NL	Natural Language
NLP	Natural Language Processing
NPL	National Physical Laboratory
OCR	Optical Character Recognition
OGI	Oregon Graduate Institute
OOV	out-of-vocabulary
OS	Operating System
PABX	Private Automatic Branch Exchange
PBX	Private Branch Exchange
PAM	Pulse Amplitude Modulation
PB	phonetically balanced
\mathbf{PCM}	Pulse Code Modulation
PD	Phonetic Discrimination
PDA	Pitch Determination Algorithm
PDI	Pitch Determination Instrument
PEB	Pulse coded modulation Expansion Bus
PIN	Personal Identification Number
PMT	Parametric Test Manager
POS	Part Of Speech
PSTN	Public Switched Telephone Network
PTS	Progiciel de Traitement de Signal
PZM	pressure zone microphone
RAM	Random Access Memory
ROARS	Robust Analytical Recognition System
ROC	Receiver Operating Characteristic
RSRE	Royal Signals Research Establishment
RT	Reverberation Time

SAM	Speech Assessment Methods (ESPRIT project)
SAMITPRO	SAM Iterative Proportional Fitting
SAMPA	Speech Assessment Methods Phonetic Alphabet
SAMPROSA	SAM PROSodic Alphabet
SAMTRA	SAM TRanscription Analysis
SAM_SCOR	SAM Input Assessment Scoring Software
SAM_SLM	SAM Speech Level Measuring program
SAM_SPEX	SAM Speech Parameter EXtractor
SCATIS	Spatially Coordinated Auditory/Tactile
	Interactive Scenario
SCSA	Signal Computing System Architecture
TAO	Telephony Application Object
SCR	System Correction Rate
SDL	Specification and Description Language
SDS	Spoken Dialogue System
SDT	Signal Detection Theory
SGML	Standard Generalized Markup Language
SU	speech input
SIFT	Simple Inverse Filtering Tracking
SIO	Speech Input/Output
SL	Spoken Language
SLM	Speech Level Meter
SLP	Speech Level Meter Spoken Language Processing
SLWG	(EAGLES) Spoken Language Working Group
SNR	Signal-to-Noise Batio
SOAP	Speech Output Assessment Package
SOS	Second-Order Statistics
SPL	Sound Prossure Level
SOL	Standard Query Language
CC CC	Shandard Query Language
SS	Speech Jynthesis Speech Technology Assessment Group
SING	Speech rechnology Assessment Group
SIIC	Somentically Unpredictable Sontences
SUS	Speech Understanding System
UUU TEI	Text Encoding Initiative
TEW	Time Frequency Warping
	Transaction Success
	Telephony Services Application
ISALI	Programming Interface
TTS	Togramming interface
	Tence and Dreak Indiana (transariation)
TODI	University College Londer
UCL	University Conege London
UUK	User Correction Kate

VR	Voice Response
VRU	Voice Response Unit
WIF	Word form Interchange Format
WORM	Write Once Read Many
WOZ	Wizard of Oz
WSJ	Wall Street Journal

 χ test, 358 μ -law, 57, 124, 294 m-gram model, 241, 243, 260, 261-262, 266-268, 277 n-gram model, 43, 395, 403, 404 ATIS corpus, 96, 103, 125, 146, 161, 169CRIL (Computer Representation of Individual Languages), 83, 84, 150, **151–152** Marsec, 150, 162, 163, 164-165, 165, 168, 172 Morpa, 514, 517 POLYPHONE corpus, 102, 103, 117, 124, 129, 156, 157, 169 SAM-A, 462 SAM, 2, 6, 7, 7, 10, 20, 150, 152, 166, 180, 208, 222, 326, 336, 381, 392, 398-401, 483, 491, 493, 501, 522, 524, 529, 542, 544, 547, 550, 595SAM Overall Quality Test, 496, 506, 550, **561** SAM Prosodic Form Test, 531, 559 SAM Prosodic Function Test, 533, 560SAM Standard Segmental Test, 503, 520, 522, **522**, 523, **523**, 524, 525, 539, 550, **551** Speechdat, 7, 155, 158, 169 Sqale, 7, 402, 405, 406 VERBMOBIL, 7, 126, 129, 146, 148, 150, 152, 153, **154–155**, 156, 157, 165-168, 178-180, 183, 193, 211, 227, 228, 599 A-law, 57, 124, 136, 294, 314, 315 accelerations and movements, 45 accent, 60, 164, 166, 167, 187, 196, $219,\ 220,\ 481,\ 482,\ 490,$ $514,\ 517,\ 525,\ 530,\ 531,$ 535, 537, 538, 538, 545, 546, 549

```
foreign, 408, 466
in MARSEC, 164
```

lexical, 60 phrase, 164, 167 pitch, 164, 166, 167, 188, 209 regional, 35, 36, 118, 158, 298, $409,\ 466,\ 584$ sentence, 100, 102, 486, 517-518syntactic, 60 tonal, 205, 210 accent identification, 409 acceptable shortfalls, 33 acceptance, 411, 414, 426, 446, 450, 451, 453, 456, 458, 462, 470acceptance threshold, 393 accuracy, 32, 182, 348, 363, 385, 388, 390, 394, 456-458, 460, 473, 519, 603 of classification, 367 of pitch determination, 112 of pronunciation, 112, 112, 508 of segmentation, 153, 154, 159, 172, 346, 364 of the MORPA module, 517 acoustic environment, see environment, acoustic acoustic interface, 486, 518-540 vs. linguistic interface, 488 acoustic module, 58, 60-61, 73, 546 acoustic testing, 547-550 acronym, 59, 82, 196, 511, 512 action, 568-570, 572, 603 Adaptive Differential PCM (AD-PCM), 57 adverse conditions, see environment, adverseaffix, 158, 186, 188, 193, 199, 212, 214, 215, 215, 516, 553 affricate, 206, 367 AGC, see automatic gain control age, see speaker, age of age identification, 51, 408, 464 age verification, 409, 413 agent, 568, 569, 569 agglutinative language, 197, 215 air-flow measurement, 131-132 alignment, 387, 388, 407

DTW, 364 forced, 191 label, 233 phoneme, 386 phonetic, 407 segmental, 500, 549 speaker, 411 time, 149, 272, 278, 279 word, 386 allophone, 93, 108, 160, 180, 206, 209, 492, 498, 499, 542 allophone synthesis, 492, 498, 499, 542alternation morphographemic, 199 morphophonological, 199, 214-216orthographic, 215, 216 amplitude distortion, see distortion, amplitude amplitude-frequency curve, 333 Analysis of Variance (ANOVA), 354, $356, \ \mathbf{356}-\mathbf{357}, \ 364, \ 375,$ **389**, 391, 401 analytic approach, 39, 486, 491-493, 507, 526, 532, 533, 537, 543 - 544vs. global approach, 490 anaphora, 218, 570, 571, 603 anechoic chamber, 95, 310, 311, 316, 322, 324, 327, 329, 331, 337, 338, 418, 600 ANN, see Artificial Neural Networks (ANNs) annotation, 4, 7, 16, 17, 22, 26, 79, 80, 86, 117, 120, 146, 147, 157, 161-163, 168, 169,171, 185, 186, 225, 546 levels of, 149, 155, 158 of dysfluencies, 157 prosodic, 84, 162, 164 segmental, 161 types of, 149 ANOVA, see Analysis of Variance (ANOVA) antonymy, 202, 202 aperiodic motion, 115, 291-292

aphasia, see pathological speech, aphasia API, see application programming interfaces (APIs) applicant speaker, see speaker, applicant application generator, 62, 64, 67, 70, 75application programming interfaces (APIs), **64**, 64, 66, 68, 70, 75application requirements, 30, 31, **32-33**, 34, 51 archiphoneme, 207, 226 Army Brat, 117 ARPA, 9, 20, 180, 402, 404-407 Articulation Index, 489articulatory parameter, 130 Artificial Neural Networks (ANNs), 35, 346, 348, 353, 360, 361, 365, 366 artificial test signals, 24, 385, 392 aspiration, 154, 160, 206 assimilation, 58, 100, 151, 159, 209, 220, 514, 517, 546 audio port, 66 audiology, 91-92 audiometry, 101, 299, 300-301 auditory environment, see environment, auditory automated testing/assessment, 485, 486, 541, **541–543** vs. human subjects, 488–489 automatic gain control, 48, 400-401automatic segmentation, see segmentation, automatic average noise consumption, 300 backing-off, 248, 249, 249-252, 252, **252–253**, 253, 254, 258, 259, 261, **281-284**, **286**-**287**, 404 badger, 441 band limiting, 384 $bandwidth, \ 34, \ 45, \ 55{-}57, \ 71, \ 73,$ 124, 133, 134, 294, 316, 333, 337, 340, 383, 384,

393, 418, 424, 426, 468, 478, 510 Bayes decision rule, 238-240, 245, 271beam search, 271, 273, 273, 274, 278 Bellcore Test, 520, 524, 524, 540, 553benchmark, 6, 20, 24, 345, 386, 391, 394, 395, 398, 402, 404-407, 491, 497, 497, 512, 522benchmark test, 74, 344, 402, 479, 483, 497, 497 between groups variance, 357, 357 bias, 139, 348-349, 360, 364, 371, 428, 443 bidirectional microphone, see microphone, bidirectional bigram, 43, 94, 182, 187, 191, 237, 241, 243, 246, 248, 249, $251 - 261, \quad 264 - 270, \quad 273, \quad$ 274, 277, 278, 395, 404 phone, 243 binaural processing, 340, 341 bit-errors, 393 black box, 24, 44, 54, 56, 59, 65, 331, 462, 484-486, 486, 487, 495, 497, **502–511**, 518, 527, 540, 566, 603, **604**, 605, 614vs. glass box, 485-487 blockage level, 69 BNC, see British National Corpus (BNC) Bode diagram, 333 bracketing paradoxes, 202 branching factor, 41, 382 breathy voice, 112, 115, 297, 297, 298, 320 British National Corpus (BNC), 121, 125bus expansion, 47 Multi-Vendor Integration Protocol (MVIP), 47, 66 PC, 65

Pulse coded modulation Expansion Bus (PEB), 66 Signal Computing Bus, 66 Signal Computing System Architecture (SCSA), 47, 66 cache model, $\mathbf{256}$ - $\mathbf{257}$ calibration, 122, 134, 315, 317, 338, 385, 390, **391**, 400, 461, 463, **463-464**, 480 call tracking, 70 call volume, 68 candid microphone speech, 120-121 canned speech/text, 56, 58, 73, 220, 481, 520, 568, 583, 592 canonical form, 83, 158, 170, 187, 188, 199, 596 canonical inflected form, 196, 199, 200, 203 canonical phonemic representation, 146, 202 canonical pronunciation, 188, 205 capability profile, 32, 32-33, 384-**385**, 396 carrier phrase/sentence, 58, 73, 100, 122, 337, 392, 394, 481, 559CART (Classification And Regression Trees), 107, 243, 263 categorical estimation method, 496, **496**, 506, **506**, 507, 540, 561, 562 CCITT, 57, 59, 573 cell, 389, 391 CFA, see Confirmatory Factor Analysis (CFA) channel, 22, 23, 36, 38, 43-44, 44-46, 50, 65, 66, 71, 73, 97, 97-98, 101, 120, 124, 127, 129, **130–132**, 133, 135, 136, 289, 290, **331**, 336, 337, 338, 338, 340, 379, $380, \ 383, \ 384, \ 393, \ 418,$ 424, 464, 466, 473, 480, 481, 489, 544, 583, 588-**589**, 591 cineradiography, 131

circuit-switching, 66 citation form, 83, 84, 129, 152, 158, 159, 170, 188, 205, 208, 209, 514 Classification And Regression Trees, see CART CLID, see CLuster IDentification (CLID) Test clipping, 335, 393 clitics, 164, 189, 220 closed response identification, 554, 555closed-set identification, 431-460CLuster IDentification (CLID) Test, 520, 523, 524, **524**, 525, 529, 539, 544, 547, 550, 552cluttering, see pathological speech, cluttering co-hyponymy, 201, 201, 202 coarticulation, 58, 87, 93, 99, 100, 244, **298–299**, 486, 499 cognitive activity, 121 cohort, 421 command system vs. interactive dialogue system, 569 command-and-control, 394 common-password speaker recognition system, see speaker recognition system, common-password communication chain, 288, 289-290 communication channel variables, see variables, communication channel communication line, 397 communication means, 601 communication modalities, 601 compactness, 554 comparative testing, 483, 486 competence (vs. performance), 90, 91, 574, 577 compound, 100, 158, 186, 195, 196, 204, 205, 210, 212, 214, 215, 227, 516

compression technique, 223

computer telephony integration, 67-**68**, 75 concept-to-speech (CTS), 58, 481, 482, **482** condenser microphone, see microphone, condenser confidence area, 52 confidence ranks, **434–435**, 436 Confirmatory Factor Analysis (CFA), 372 confusion matrix, 42, 43, 53, 71, 388, 391, 401, 434, 437, 527, 551, 552, 598 confusions, 41, 42, 52, 388, 392, 526 cross-sex, 420same-sex, 420connected words/speech, 37, 40, 50, 62, 83, 91, 92, 97, 99, 100, 117, 159, 272, 381, 382, 386, **387–388**, 392, 394, 395, 397, 399 consonant cluster, 108, 524-526, 552 constraints language model, 240, 247, 273 phonotactic, 247, 524, 525, 552 pragmatic, 240, 241 semantic, 240 syntactic, 240 contemporaneous speech, 419 context free, see grammar, context free contextual appropriateness, 605.**606-608**, 614 continuous speech, 24, 37, 44, 83, 95, 96, 178, 180-182, 185, 239, 270, 272, 381-384, 387-**388**, 394–396, **401–601** vs. isolated words, 95 controlled experiment, 90, 106, 360 cooperative speaker, see speaker, cooperative cooperativeness, 73 copyright, 85, 404 corpus perplexity, 237, 245, 246, 248, 260correction rate, 25, 605, 608-610 coverage, 189, 190, 246, 251, 257

bigram/trigram, 246 vocabulary, 246, 251 creaky voice, 297, 298 cross-language comparability, 491, 501 - 502cross-validation, 237, 248-250, 260 crosstalk, 384 CTS, see concept-to-speech (CTS) cut-through facility, 48-49, 69 CVC-words, 384, 391, 392, 394 damping distortion, see distortion, damping DARPA, 6, 181, 402, 595 data model, 192, 193 database management, 81, 565 database management system (DBMS), 138, 222, 224, 401DATR, 180, 194, 211, 223 decibel scale, 292 declarative knowledge base, 184 default inheritance hierarchy, 192 deletion, 88, 114, 154, 157, 159, 169, 170, 180, 181, 209, 220, 226, 281, 367, 369, 386-388, 407, 515, 525 demisyllable, 214, 239, 361, 499 dependable speaker, see speaker, dependable dependent variable, see variable, dependent derivation, 196, 210, 212, 213, 214, 215derivational affixation, 215 design by intuition, 573-578 by observation, 578-581 by simulation, 581-591development test, 398, 402, 405, 405 devoicing, 160, 207 diagnostic approach, 18, 24, 384- $386, \ 391, \ \textbf{391} \textbf{--392}, \ 394,$ 405, 406, 420, 428, 444, 451, 483, 484, 486, 487, 493-495, 497, 518, 520-522, 524-527, 529, 532,

537, 538, 542, 543, 547, 549, 551-555, 559, 560diagnostic corpora, 18, 372-373 Diagnostic Pairs Sentence Intelligibility Evaluation Test (DPSIE), **525**, 526 Diagnostic Rhyme Test (DRT), 392, 463, 494, 520, **520**, 521, **521**, 522, 525, 547, **554** dialect, 73, 94, 116-118, 124, 156, 171, 211, 219, 390, 392- $394,\ 424,\ 494,\ 535,\ 538,$ 579, 584dialectology, 88, 116, 146 dialogue menu-driven, 41, 62, 570-571 transcription of, 148, 170 dialogue duration, 605, 605, 614 dialogue grammar, 572, 595, 598 dialogue history, 570, 571 dialogue manager, 41, 62, 62, 63, 568-571, 579, 587, 603 dialogue model variables, see variables, dialogue model dialogue strategies, 567, 585, 598 adaptive, 598 constitutive, 598 cooperative, 598 deterministic, 598 dialogue system, 30, 58, 157, 182, 262, 281, 344, 347, 482, 564 - 615automatic, 238 interactive, 15, 16, 24-25, 178, 375 - 380, 564, 565, 567, **567–571**, 572, 580, 594, 595, 600, 603, 604, 613, 614vs. command system, 569 menu, 41, 570-571, 576, 581 multimodal, 571 spoken language, 4, 15, 62, 189, 220, **571**, 572, 577, 578, 580, 581, 590, **591–594**, 594, 595, 608, 614 dictation, 5, 27, 37, 121, 127, 178, 187, 219, 238, 246, 247,

257, 262, 275, 382, 394, 404dictionary, 23, 38, 41, 53, 56, 60, 155, 187, 188, 190, 217, 247, $402,\;403,\;407,\;514,\;516$ Duden, 156, 179, 182 exceptions, 515, 516 of phonetic elements, 99, 108 on-line pronouncing, 158 Oxford English Dictionary, 179 phonemic, 158 phonetic, 59 pronouncing, 514, 545 pronunciation, 179-182, 239, 240Digital Audio Tape (DAT), 293, 314, 317digital signal, 47, 47, 56, 66, 293-**295**, 313–317, 334, 335, 337, 339, 397, 399-401, 470dimensions of difficulty, 32 diphone, 60, 74, 93, 93, 101, 102, $108, \ 186, \ 214, \ 233, \ 243,$ 361, 362, 486, 492, 499, **523–524**, 534–536, 538 Diphone Test, 520, 523-524, 524 diphthong, 147, 154, 206 disambiguation, 177, 487, 517, 533, 534discounting absolute, 237, 249, 252-253, 254-256, 258-261, 286-287 linear, 237, 248, 249-252, 252, 253, 255, 257, 260, 261,281-284, 285, 286 discourse analysis, 91, 104 distortion, 34, 45, 46, 71, 91, 114, 115, 294, 306, 312-314, 316, 331, 332, **333–335**, $478,\;480,\;489,\;500$ amplitude, 333-334damping, 333 linear, 288, 309, 333-335, 335 non-linear, 288, 314, 335, 335

phase, 134, 314, 317, 333, 334-335 disyllable, 502 documentation, 79, 156, 157 domain, 383 domain expertise, 584, 600 Doug Paul, 404 DPSIE, see Diagnostic Pairs Sentence Intelligibility Evaluation Test drivers, 63, 64, 75 drop-out, 384 DRT, see Diagnostic Rhyme Test (DRT)dry run, 402, 405 DTMF, 46, 47, 57, 61, 63, 470, 568 dummy head, 324, 325-326 dummy item, 128 duration, 55, 106, 151, 154, 162, 163, $171,\ 359,\ 365,\ 379,\ 531,$ 543, 614 dialogue, 69, 126, 605, 605, 614 mean vowel, 354 of a call, 68, 69 of a sentence, 243 of average noise consumption, 300 of recording, 125, 142 of training phase, 36, 38, 40 phoneme, 180, 499, 534 rule-derived, 531 segment, 163, 380, 499, 519,531, 548 syllable, 298, 299, 515 tone unit, 355, 356 turn, 605, 606, 614 duration model, 60 duration rules, 151, 499, 531, 534 duration structure, 499 dynamic microphone, see microphone, dynamic dynamic programming, 272-275, 277, 278, 280 dynamic range, 124, 133, 384, 399 Dynamic Time Warping (DTW), 35, 364, 462, 463, 542 dysarthria, 114

dysphonia, 115 E1 (USA), 47, 67 ear-training, 492 echo, 45, 46, 48, 71, 135, 171, 314, 384, 393 echo cancelling functions, 48 EER, see equal error rate (EER) efficiency, 333 electrical characteristics, 384 electroglottogram, 108 electromagnetic articulography (EMA), **131** electromyograph activity (EMG activity), 108 electromyograph signal (EMG signal), 79, 122, 130 electropalatography, 22, 130, 130-**131**, 151 EM algorithm, 237, 252, 262, 267, 285emotion, 51, 73, 105, 111, 115, 400, 466, 530, 533-535, 549 emotion-by-rule, 534, 535 emulation packages, 67 energy threshold, 50, 391, 393 environment, 8, 19, 22, 23, 34-36, 38, **43-44**, 45, 70, 71, 95, 115, 119-122, **122**-125, 133, 135, 137, 288-290, 301, 306-310, 321-**331**, 336–338, 340, **341**, 424, 464, 466, 468, 480, 488, 498, 503, 511, 569, 571, 597, 600, 601, 602, 613 acoustic, 55, 110, 122, 290, 306, 309, 310, 341, 393, 600 adverse, 50, 137, 384, 600 anechoic chamber, see anechoic chamber artificial vs. natural, 321-322 auditory, 288-290, 322-330, 341development, 58, 61, 68 familiar, 99 field, see field

in-car, 290, 315 laboratory, see laboratory living, 538 noisy, see noise open, 66 recommendations on recording environment, 316 reverberant, see reverberation speech assessment, 290-291studio, see studio test, 602 very quiet, 311, 315, 329, 331 visual, 288-290, 330-331 environmental control device, 115 epenthesis, 209 equal error rate (EER), 451-460, 469, 471 equal loudness contours, 293 equalisation techniques diffuse-field, see headphone, diffuse-field equalised free-field, see headphone, freefield equalisation Erlang, 69, 70 $error \ rate, \ 72, \ 129, \ 245, \ 271, \ 385-$ 388, 570, 582 equal, see equal error rate (EER)sentence, 386, 390 word, 386, 388, 390 ESPRIT, 1, 6, 7, 10, 20, 93, 108, 180, 208, 327, 330, 381, 398, 483, 491, 514, 522, 524, 550estimator, 349, 386, 387 ${\rm EUROM\text{-}1},\ 7,\ 21,\ 83,\ 96,\ 108,\ 130,$ 392event-dependent speaker recognition system, see speaker recognition system, eventdependent exchange, 568, 569, 573, 603, 604 expectation-maximisation algorithm, see EM algorithm exploitation phase, 35, 53 Exploratory Factor Analysis, 372 external NLP module, 43

F ratio, 354 face-to-face spoken dialogue translation, 482, 599 factor, 348, 357, 357, 371, 372 factorial experiment, 106-107, 108, 109 false acceptance, 49, 51-53, 412, 414, 436, 440-444, 444-453, 455, 457-460, 468, 470-472, 475, 476 false alarm, 368, 387, 387, 388 false rejection, 51-53, 412, 414, 436, **439-440**, 444-447, 450-453, 455, 456, 458-460, 468, 470, 472, 475, 476 FFT data, 80, 119 field, 5, 41, 71, 109, 123, 144, 389, 423, 574 field testing, 19, 24, 41, 42, 74, 291, 385, 389, 469, 470, 472, 475, 477, 484-486, 507, **508-511**, 536, 540, **544**-**545**, 592, 593, 602 vs. laboratory testing, 487 -488 figure of merit, 387, 388, 461, 484 filter bank data, 80, 119 finite state automaton, 61 finite state grammar, see grammar, finite state finite state language model, see language model, finite state finite state network, 238, 241, 243, 267, 268, 274 finite state syntax, see syntax, finite state first language acquisition, 89 fixed-vocabulary speaker recognition system, see speaker recognition system, fixedvocabularv flawless speech, 290, 290, 291, 301, 301, 334, 340 flowchart, 572, 594 fluency, 37, 71, 115, 393, 508, 509 focus, 166, 167, 482, 515, 530, 534,

603

foot, 162, 163, 210 formant extraction, 80, 110, 112, 119 frequency distribution, 107, 478 frequency response, 134, 293, 302, $304,\ 306,\ 312,\ 324,\ 333,$ 333, 334, 335frequency spectrum, 292, 293, 297, 309, 324, 333-335, 340 fricative, 124, 147, 153, 160, 367 Friedman test, 358 full trigram model, 237, 254-255 functional disorder, 114 functional testing, 486, 490, 491, 495, **502–506**, 506, 507, 509, 511, 520, **520**, 526, 527, 530, **533–535**, 538, **543**, 548 vs. judgment testing, 489-490 fundamental frequency (F0), 124, 161, 162, 164, 165, 168, $209, \quad 210, \quad 288, \quad 291{-}294, \quad$ 296-298, **299**, 300, 302- $304, \quad 306{-}310, \quad 312{-}315, \quad$ 318, 319, 324, 332-335, 337-340, 348, 391, **392** Fuzzy Logic, 190 G-effect, 600 gating experiment, 88 gender, 406 gender ratio, 390, 406 gender-balanced, 432, 434-436, 439, 442, 443, 445, 453, 454, 457 - 461general public, 35, 102, 474, 493, 494, 571, 586, 600 Generalised LInear Modelling Techniques (GLIM), 375 genuine speaker, see speaker, genuine glass box, 24, 59, 484-486, 486, 487, 495, 497, **511–540**, 602, **603**, 614 vs. black box, 485-487 GLIM, see Generalised LInear Modelling Techniques (GLIM)

global approach, 39, 486, 493, 537, 543 - 544vs. analytic approach, 490 glottal, 134, 160, 289, 297, 318-320, 338 glottal stop, 160 glottal-to-noise excitation, 296, 298 goat, 43, 394, 432, 476 grammar, 2, 43, 118, 180, 181, 216- $\mathbf{219}, \ \ 219, \ \ 237, \ \ 238, \ \ 243,$ **266–269**, 381, 382, 384, 395, 403, 575, 578, 590, 594, 595 *n*-gram, see *n*-gram model bigram, see bigram competence, 91 context free, 243, 266, 267, 269, 276dialogue, see dialogue grammar finite state, 180, 267, 268 lexicalised, 267 link, 237, 267, 270 null, 181, 182 phrase structure, 243 probabilistic, 253, 266, 395, 403, see language model, probabilistic sentence, 390 stochastic, 243, 266, 276 traditional school, 192 word-pair, 181, 182, 395 grammar checking, 517, 545 grapheme-to-phoneme conversion, 84, 101, 146, 155, 158,170, 178, 204, **513–515**, 516, 545, 546 Graphical User Interfaces (GUIs), 64 graveness, 554 greedy algorithm, 101, 107, 108 GUI, see Graphical User Interfaces (GUIs) gullibility, 585 handheld microphone, see micro-

handheid microphone, see microphone, handheid harsh voice, 297, **297**, 298, 535 Harvard Psychoacoustic Sentences, **527–528**, 528, 529, **558** Haskins Syntactic Sentences, 528, 528 - 529, 529, 556HCI, see human-computer interaction (HCI) HCI (human-computer interface), 16 HDSCAL, 401 headmounted microphone, see microphone, headmounted headphone, 19, 290, 306, 316, 323, **323–327**, 328–330, 334, 510diffuse-field equalised, 324-**326**, 329 free-field equalised, 324, 325, 329open vs. closed, 323-324, 329 headsetmicrophone, see microphone, headset health, 36, 51, 98, 131, 141, 144, 392, 393, 408, 464 health state identification, 51, 408 height, see speaker, weight/height of heterography, 201, 201 heterophony, 201, 201, 205 Hidden Markov Model (HMM), 35, 41, 42, 94, 151, 153, 177, 181, 187, 190, 224, 239, 271, 275, 362, 393, 417, 578 histogram, 207, 246, 350, 562 HMM, see Hidden Markov Model (HMM) hoarseness, see pathological speech, hoarseness homography, 185, 201, 203, 205, 514 homonymy, 201, 201 homophony, 201, 201, 225, 519 host-computer, 67, 75 HPSG (Head-Driven Phrase Structure Grammar), 193 human subjects automated VS. assessment, 488 - 489

human-computer interaction (HCI), 18, 104, 579-582, 587, 593, 606, see asoman-machine interaction (MMI)857 human-computer interface, 4, 5, 16, 20, 487, 533, 548, 560 human-human interaction, 579-581, 590, 593, 599, 604 hyperbolic model, 455 hyperonymy, 201, 201, 202 hyponymy, 201 hypothesis testing, 354-360 identification, see spoken language identification identity assignment, 411, 413, 461 idiolect, 156, 579 idiom, 196, 220-221 phrasal, 193, 196, 205 pragmatic, 189, 220 idiosyncrasy, 107, 193-195, 205, 210, 515ILEX (Integrated LEXicon, 180, 211 imitation ability, 444-445 impostor, 51-53, 143, 411, 413, 429-431, 469, 471, 475, 476 acquainted, 422, 423, 427, 442 casual, 422, 423, 425, 442, 446 intentional, 422, 423, 425, 427, 442, 464 poor, 441, 449 pseudo-, 421, 425, 451 skilled, 441, 442, 449 unacquainted, 422, 423 unintentional, 464 well-intentioned, 422, 423, 425 impostor model, 53, 421, 427 in-situ recording, **41**, 71 inbound calls, 47 independent t test, 358, 359 independent variable, see variable, independent infix, 212 inflection, 179, 182, 183, 185-187, 190, 191, 193, 196, 197, **197–198**, 199, 200, 202,

203, 208, 212-214, **214**, $215, \ 216, \ 227$ information slots, 58, 481 insertion, 260, 281, 367, 369, 386-388, 407, 525, 608 insertion-loss compensation, 326-**327**, 328, 329 intensity, 79, 112, 116, 161, 293, 299, 300, 337-339, 515, 549interaction, 389, 568 Interactive Voice Response (IVR), 67, 570, **591–592**, 598 interactive voice system, 21, 61-63, 570, 577, **591–592**, 598 interjection, 157, 186, 218 internal switching matrix, 66 International Phonetic Alphabet, see IPA International Phonetic Association, 151, 208 interoperability, 66 interpolation, 237, 248, 250, 252, 254, 255, 257-259, 262, 270, **284-286** interpreting telephony, 482, 482. 536intonation, 60, 106, 108, 115, 151. 161, 162, 164-168, 210, 211, 217, 219, 220, 297, 299, 318, 493, 530, 532, 533, 546, 559, 560, 580, 586, 588 intonation pattern, 165 IPA, 83, 152, 158, 159, 172, 183, 203, 208, 209, 211, 226 IPO, 162, 163, 165, 168, 172 ISDN, 47, 124, 136, 313-316, 470, 473isolated phonemes, 92, 99 isolated sentences, 91, 100-102, 102, 518, 531, 532 isolated words, 24, 37, 42, 50, 79, 83, **95**, 95, 96, **99–100**, 108, 109, 120, 381, 386-**387**, 388, 394–396, 401, 407, 514, 547, 570, 601

vs. continuous speech, 95 $isophone,\ 293,\ 338$ ITU, 507 ITU-T Overall Quality Test, 507, 508, 536, **562** IVR, see Interactive Voice Response (IVR) jitter, 296, 298, **298** Jonkheere's trend test, 358 judgment testing, 486, 490, 494-496, 509, 511, 520, 527, 532, 533, 537, 538, 541, **543**, 548at the word level, **526** laboratory, 506-508of prosody, 531-533 vs. functional testing, 489-490 Kendall's W, 358 keyword, 32, 40, 61, 70, 73, 527-529, 556, 558 knowledge-based systems vs. stochastic systems, 94 Kruskaal-Wallis, 358 labelling, 22, 35, 36, 38-41, 44, 54, 60, 80, 81, 120, 142, 146, 147, 150-155, 158-161, **163–168**, 168, 172, 200, 225, 233, 318, 340, 345, 346, 360, 363-365, 367, 373, 411, 513, 516, 595in ToBI, 163-164 in Marsec, 164-165 in Verbmobil, 154-155 prosodic, 161-162, 163, 168 in Verbmobil, $\mathbf{165}\text{--}\mathbf{168}$ speaker, 411, 466 laboratory, 5, 79, 95, 109, 119, 134, 136, 302, 309, 425, 476, 578laboratory prototype, see prototype laboratory speech, 83 laboratory testing, 19, 24, 42, 74, 291, 422, 486, **502–508**,

508-511, 540, 541, **544**- $\mathbf{545},\ 592,\ 593,\ 602$ vs. field testing, 487-488lamb, 433Language Engineering Standards (LES), 346-347 language identification, 409, 464 language model, 34, 43, 71, 184, 190, 191, 213, **237–287**, 382, 402, **403-404**, 580, 591, $597, \ 601$ *m*-gram, see *m*-gram model bigram, see bigram finite state, 243, 268 grammar based, 237, 238, 243, 266-269 probabilistic, 382, 403 statistical, 216-234 stochastic, 177, 185, 187, 212, 216, 217, 231, 237, 240-243, 245, 253 trigram, see trigram types of, **243** unigram, see unigram language model constraints, see constraints, language model language model robustness, 266 laryngograph, 296, 319, 321 laryngograph recording, 130, 130 laryngograph signal, 80, 119, 130, 296, 297, 318 laryngography, 22 lattice of lexicon entries, 44 leaving-one-out, 237, 249-251, 253, 255, 260, 282, 284, 286 lemma, 193, 196, 199, 200, 203 lemmatisation, 179 LES, see Language Engineering Standards (LES) lexical access, 177 lexical database, 25, 177, 178-180, 183, 183, 184, 184-186, $187, \ 191, \ 192, \ 197, \ 198,$ $200,\ 202,\ 203,\ 211,\ 216,$ **223–230**, 231, 232, 234, 235, 516, 517 lexical information model, 192

lexical noise, 185 lexical unit, 84, 195-200, 200, 219 lexicalised grammar, see grammar, lexicalised lexicon BRULEX psycholinguistic, 179 bilingual, 90 declarative, 200, 231 EVAR, 182-183 exceptions, 513 fully inflected form, 197-198 hierarchical, 195 in SL recognition systems, 190 - 191mental, 88, 89 morph, 198, 199 morpheme, 517 procedural, 200 pronunciation, 271, 273 prosodic, 210 spoken language, 176-236 spoken language vs. written language, 187–189 stem. 199 lexicon architecture, 177, 180, 183, 191, 193, 221, 223 lexicon construction, 81, 233, 234- $\mathbf{235}$ lexicon model, 192-193, 202, 203, 223, 231, 235 lexicon size, 38, 71, 198, 208 lexicon structure, 221–231 lexicon theory, 192, 193 lexicon-entry identifier, 44 liaison, 58, 159, 226 Linear and Time-Invariant (LTI) systems, 332 linear distortion, see distortion, linear linear interpolation, 237, 248, 252, 257, 259, 262, 270, 284-286 linguistic interface, 482, 486, 511-518, 541, 545-546 vs. acoustic interface, 488 linguistic module, 44, 58-60, 488, 511, 545

lingware, 223, 235 link grammar, see grammar, link listener descriptors, 299-301 listening effort, 61, 501, 507, 508, 562 ${\rm Lombard} \ {\rm effect}, \ 83, \ 98, \ 383, \ 384,$ 392, 464 LPC data, 80, 93, 119, 161, 498 LTI system, see Linear and Time-Invariant (LTI) system magnitude estimation method, 373, 496, 496, 506, 506, 507, 540, 560, 561 man-machine interaction (MMI), 30, 31, 96, 105 Mann-Whitney U test, 358, 360 manual segmentation, see segmentation, manual McNemar test, 359, 401 Median Test, 358 menu-driven system, 61-62 meta-textuals, 512 microphone, 19, 22, 36, 43, 44, 45-46, 46, 50, 55-57, 65, 70, 79, 95, 98, 104, 110, 119, **120–121**, 121–123, 125, 133, 134-135, 135, 144, 288, 290, 296, 301, 302-**306**, 306, 308-311, **311**-**312**, 313-315, **315-316**, 316, 317, **317**, 317, 318, 320, 324, 326, 327, 329, 331, 337, 338, 384, 393, 397, 400, 418, 424, 426, 427, 466, 600 bidirectional, 305 close-up, 135 condenser, 122, 134, 302-303, 315, 316 dynamic, 134, 302 handheld. 45 headmounted, 45 headset, 134, 135, 302, 306, 316, 317, 327, 329 omnidirectional, 134, 303, 305, 306, 315, 317

positioning of the, 45, 123, 134, 135, 311, 315, 316, **317**, 329press-to-talk, 45, 70 pressure zone, 306, 312 recommendations on, 315-316 room, 135 table-top, 135 ultradirectional, 305 unidirectional, 134, 303, 303-305, 305, 315, 317 cardioid, 304, 305, 315 hypercardioid, 304, 305 supercardioid, 304, 305 microphone array, 45, 104, 134, 600 microphone installation, 45 microphone interfaces, 45-46 Minimal Pairs Intelligibility Test (MPI), 520, **525**, 526 misclassification, see speaker misclassification mistaken speaker, see speaker, mistakenmistrust, 431, 432-434, 435-437, 459MMI, see man-machine interaction (MMI) MNRU, see Modulated Noise Reference Unit (MNRU) model generation, 40 Modified Rhyme Test (MRT), 494, 520, **520**, 521, **521**, 522, 522, 539, 540, 547, 555 Modulated Noise Reference Unit (MNRU), 500 monitoring, 120, 129, 475, 599 on-line, 121, 122 phoneme, 88, 490 svllable, 490 word, 490, 505, 540 monologue, 18, 103, 148, 164 mood identification, 51, 408 morph, 186, 186, 193, 213, 214, 215 bound, 215 free, 215 lexical, 215 orthographic, 199, 215

phonological, 199, 215 morph boundary, 205, 213, 215 morpheme, 59, 186, 187, 193, 199, **210**, **212**, 213, 214, **214**, 215, 226, 227, 515, 516, 534grammatical, 215, 215 lexical, 196, 199, 214, 215 morpheme boundary, 100, 516 morphographemic alternation. seealternation, morphographemic morphological decomposition, 59, 199, 213, 514, 515, 516-517morphological generator, 215 morphology, 188, 191, 194, 197-199, 202, 205, 207-210, **212**- $\mathbf{216},\ 217,\ 223,\ 226$ applications of, 213-216 inflectional, 187, 198 recommendations on, 216 two-level, 198, 216 morphophonemics, 194, 207, 208, 208morphophonological alternation. see alternation, morphophonological morphophonological rules, 208, 216 morphotactic rules, 208, 215 morphotactics, 214, 217 MPI, see Minimal Pairs Intelligibility Test (MPI) MRT, see Modified Rhyme Test (MRT) MS-TAPI, 68 multi-channel system, 50, 65, 66 multi-layered hardware and software architecture, 66 multi-speaker system, 35 Multi-Vendor Integration Protocol (MVIP), 47, 66 multiplexed digital telephony highway, 66 MVIP, see Multi-Vendor Integration Protocol (MVIP)

nasal, 114, 131, 153, 154, 160, 170, 205, 412 nasalisation, 154, 160 nasality, 537, 554 National Institute of Standards (NIST), 6, 14, 20, 402, 407 NATO, 8, 156, 400 natural language processing (NLP), 43, 81, 82, 86, 96, 177, 189, 216, 222, 223, 574, 578,579naturalness, 59-61, 74, 99, 102, 103, 121, 161, 171, 490, 506, **506**, 508, 526, 527, 531, 559, 561 Neural Network, 35, 94, 190, 346, 348, 353, 360, 361, 365, 366newspaper, 81, 101, 102, 108, 139, 383, 394, 403-405, 492, 494, 504, 508, 509, 512-515, 517, 518, 536, 537, 545NIST, see National Institute of Standards (NIST) node, 268, 280, 281, 395, 598 noise, 8, 24, 34, 36, 41, 43-45, 48, 48, 62, 70, 72, 73, 83, 88, 95, 98, 121, 123, 128, 129, 135, 137, 157, 161, 168, 170, 171, 185, 190, 226, 253, 292, 294, 298, 300, 303, 304, 306, 310-313, $315, \ 316, \ 318, \ 322, \ 327,$ 329, 335, 337, 338, 367-369, 373, 383, 384, 393, $399, \ 401, \ 424, \ 427, \ 464,$ 470, 477, 478, 488, 489, 510, 541, 544, 573, 574, 600 addition of, 294, 390, 397, 398, **399–400**, 521, 522, 540, 550background stationary, 45 equivalent input noise (EIN), 312frication, 292

lexical, see lexical noise line, 46, see line noise orthographic, see orthographic noisepink, see pink noise pop, see pop noise signal-to-noise ratio, see signalto-noise ratio white, see white noise noise rejection, 305, 306, 393 Noisex, 400 non-linear distortion, see distortion, non-linear non-linearities, 46, 302, 335, 384 non-registered speaker, see speaker, non-registered non-speech sounds, 37 nonsense word, 99, 100, 108, 548 nuclear magnetic resonance imaging, 132null hypothesis, 355-357, 364 OCR, see Optical Character Recognition (OCR) off-diagonal elements, 388 off-line comprehension tests, 496 off-line identification tests, 495 Oglesby's model, 454-456, 458 omnidirectional microphone, see microphone, omnidirectional on-line comprehension tests, 496 on-line identification tests, 496 on-line monitoring, see monitoring, on-line one-tailed hypothesis, 355 onset, 160, 503, 548, 553 open response identification, 509, 551-553, 555-558 open-set speaker identification, 425, see speaker identification, open-set operating systems (OS), 63, 75, 231 operational capabilities, 16, 31 Character Optical Recognition (OCR), 216, 243 orthographic alternation, see alternation, orthographic
orthographic break, 213 orthographic noise, 203 out-of-vocabulary, 177 out-of-vocabulary word, 40, 42, 129, $176, \ 185, \ 246, \ 248, \ 257,$ 387, 389, 390, 393, 406 out-of-vocabulary words, 185 overall quality, 74, 490, 500-501, 506, 507, 508, 509, 519, 532, 536, 541, **550**, 550, 561, 562 PABX, 46, 48, 67, 68 Page's L test, 358 paired comparison, 373, 506, 540 PAM, see Pulse Amplitude Modulation (PAM) paradigmatic relation, 200, 201, 217, 218paradigmatic similarity, 218 parametric coding strategies, 295 Parametric Test Manager, 544 parse tree, 266, 267 parsing, 39, 43, 44, 62, 187, 217, 266, 267, 269, 270, 565, 566, 569, 571, 578, 603, 608 morphological, 215, 517 prosodic, 210, 217 sentence, 212, 217, 218, 545 syntactic, 517, 517, 545, 546 parts of speech (POS), 218, 262, 270, 501party, 568, 569 pathological speech, 86, 114, 114-**116**, 144 aphasia, 114, 115 cluttering, 115 hoarseness, 92, 114, 115, 123, 297, 298, 537stuttering, 115, 370 PCM, see Pulse Code Modulation (PCM) PDI, see Pitch Determination Instruments (PDI) performance (vs. competence), 90, 91, 573, 574, 577

 $period,\ 291,\ \textbf{291},\ 292,\ 298,\ 338,\ 340$

periodic motion, 291-292 perplexity, 23, 41, 71, 181, 182,237, **245–248**, 256, 258– 261, 265, 270, 274, 382, 390, 391, 393, 395, 406 formal definition of, 245-246 Personal Computer, 7, 48, 63, 65, 66, 68, 339, 397, 398, 401, 470 personal-password speaker recognition system, see speaker recognition system, personal-password phase, 291, 297, 312, 314, 317, 326, 333, **334-335** phase distortion, see distortion, phase phone, 93, 180, 181, 206, 206, 243, 362, 365, 367, 368, 370, 382, 386, 402, **402**, 403, 406, 499 phone model, 181, 182, 362, 403, 406phoneme, 61, 85, 88, 92, 99-101, 107, 127, 154, 158, 159, 170, 177, 180, 186, 187, 205, 205, 206, 206, 207, 209, 212-215, 217, 226.233, 239, 243, 247, 270, 271, 273, 275, 345, 347, 348, 360-363, 365, 367-369, 373, 385, 386, 402, 417, 418, 486, 490, 495, 499, **513–515**, 519, 520, 526, 530, 534, 539, 540, 547, 550, 552, 553, 601 phoneme frequency, 101, 102, 107 phoneme model, 417 phoneme monitoring, see monitoring, phoneme phonemic structure, 102, 204 Phonetic Discrimination, 392 phonetic module, 58 phonetic spelling, 181 phonetically balanced, 7, 38, 39, 53, 54, 101, 101, 361, 495, 558 phonetically rich, 101, 102 phonetics, 17, 87, 120, 146, 159, 172, 194, 492, 518

articulatory, 108 experimental, 91 pink noise, 337 pitch, 79, 80, 93, 106-108, 110, 112, $116,\ 119,\ 120,\ 124,\ 162,$ $164{-}168, \ 188, \ 209, \ 296,$ 301, 313, **318–320**, 321, 340, 379, 380, 392, **392**, 393, 482, 499, 500, 515, 517, 531, 535, 549 Pitch Determination Instruments (PDIs), 296, **318–320**, 321electrical, 319, 320, 321 mechanical, $\mathbf{318}$ photoelectric, 320 ultrasonic, **319–320**, 320 pitch movement, 164, 165, 168, 380, 515, 517, 531, 549 platform, 8, 21, 39, 53, 63-67, 68, 75, 158, 469, 572 data collection, 35, 38 development, 54 PC, 63, 65, 470 playback, 49, 56-58, 73, 137, 339, 396 - 399pleasantness, 61, 334, 490, 507-509, 526, 534, 536-538, 549, 563, 604 pooling, 255-256 pop noise, **311**, 316 portability, 1, 66, 136, 137, 151, 172, 231, 469, 566 POS, see parts of speech pragmatics, 23, 148-150, 158, 182, 184, 189, 192, 194, **219**-**220**, 220, 223, 225, 233, 239-241, 548, 570, 586, 598 pre-processing, 35, 36, 59, 79, 234, 461, 462, **511–513**, 551 pre-production parameters, 383 predictive model of performance, 30, 31prefix, 165, 196, 210, 212, 214, 215 press-to-talk microphone, see microphone, press-to-talk

private switch, 46, 48 prompt, 30, 48, 49, 56, 57, 67, 68, 81, 97, 102, 113, 126, $127{-}128,\ 128,\ 135,\ 140,$ $142,\ 330,\ 331,\ 413,\ 415-$ 418, 462, 471, 474, 576, 578, 591, 592, **592** proprietary system, 63, 65 prosody, 17, 23, 59, 60, 74, 83, 84, 93, 102, 108, 121, 126, 127, 129, 148-150, 155, 158, **161–168**, 171, 172, 187, 188, 191, 196, 200, 205, **209–211**, 211, 212, 214, 215, 217, 219, 220,226,227, 316, 340, 379, 380, 384, 417, 481, 485, 486, 499, 499, 503, 517, 518, 520, 527, 529, **530–535**, 535, 540, 541, 543, 544, 546, 547, **548–549**, 549, 550functional tests of, 533-535 functions of, 530judgment tests of, 531-533 prototype, 20, 32, 34, 67, 75, 256, 281, 475, 528, 571, 587 proximity effect, 303, 304, 304, 305, 310pseudo-impostor bundle, 421, 423, 425, 427 PSOLA algorithm, 93, 340, 482, 500, 535, 547 PSTN, 46 psycholinguistics, 80, 88-89, 90, 92, 146, 236, 495, 505, 534, 548.549 PTT approval, 48, 72 public network, 46, 48 Amplitude Pulse Modulation (PAM), 294, **294** Pulse Code Modulation (PCM), 57 linear, 294, 317, 337 logarithmic, 294 Pulse coded modulation Expansion Bus (PEB), 47, 66 pulse detection, 46, 47, 61, 63

punctuation, 59, 73, 84, 100, 102, 196, 247, 511, 533 quantisation, 85, 130, 131, 136, 137, 293, 294-295, 296 ram, 433 random sampling, 107, 110, 348, 349, 352, 362, 363 range effects, 365-366 rank order, 373, 522, 539, 540 Rapid prototyping, 67 re-prompting, 592 read speach, 359 read speech, 18, 22, 37, 38, 71, 82, 83, 99, **99–102**, 103, 109, 113, 117, 120, 125, 127, 129, 146, 147, 147-148, 149, 155, 156, 157, 163, 164, 169, 169, 171, 346, 354-356, 359, 360, 373, 383, 384, 415, 531, 578, 598omissions in, 169 transcription of, 147-148 real-time, 36, 50, 57, 65, 66-67, 72, 75, 314, 327, 384, 396, 469, 476Receiver Operating Characteristic (ROC), 369, 452-460 recognition system, see speech recognition system Recommendation G711, 57 Recommendation G721, 57 recording conditions, 384 recording phase, 81 post-recording, 81 pre-recording, 81 recording room, 306, 309-310, 316, 331recordings multi-channel, 22, 120, 127, 130-132, 331 random, 120, 125, 125, 126 $surreptitious, \quad 97, \quad 105, \quad 120, \quad$ 120-121, 144, 145, 295 visible, 120, 120, 121, 121-122 reduced word forms, 155-156, 170 redundancy, 237, 240, 246 in the lexicon, 182, 192, 195, 209, 223, 224prosodic, 533, 534 situational, 487 redundancy rules, 192, 223 reference conditions, 491, 497-502, 531, 532 melodic, 499-500 overall quality, 500-501 prosodic, 499 segmental, 498-499 temporal, 499 voice characteristics, $\mathbf{500}$ reference system, 326, 385, 390, 461-463, 497 register, 117, 219, 579, 579, 586 registered speaker, see speaker, registered rejection, 37, 41-42, 48, 49, 53, 71, 356, 369, 411, 414, 421, 423, 439, 446, 462, 469, 470false, see false rejection noise, see noise rejection rejection capability, 41, 42, 48, 395 rejection model, 42, 53 related t test, 358, 359 relative unreliability, 444-445 replica, 406 representative database, 385, 389-**390**, 405 reproducibility, 23, 288, 291, 336-338, 384, 385, 397, 417, 461, 464 requirement profile, 32, 32-33, 70-74, 384-385 RESAM, 339, 400 resistant speaker, see speaker, resis- tant response modality, 494, 495, 495, 496, **496** response time, 50, 62, 66, 72, 371, 372, 377, 378, 387, 586, 587, 588, 604

reverberation, 45, 70, 73, 122, 305, 307, 308, 310, 315, 325, $329, \ 335, \ 341, \ 384, \ 393,$ 544rhythm, 530 ROC, see Receiver Operating Characteristic (ROC) root, 165, 186, 193, 215, 215 RSG10, 8, 400 S0, 47 SAM SCORE, 7, 401 SAMITPRO, 401 SAMPA, 154, 158, 194, 203, 206, 208, 209, 211, 226 SAMPAC, 398 sampling, 294 sampling frequency, 47, 56, 112, 136, 314, 426 sampling interval, 294 sampling rate, 85, 112, 130-132, 136, 137, 294, 317, 334, 337, 340, 400 scallability, 66 scenario, 58, 69, 133, 290, 322, 329, 347, 354, 510, 565, 590, 598, 599, 601, 610-612 schwa, 99, 159, 205, 499 scoring, 386-388, 428-461 isolated word, 386-387 recognition score, 386-388 script language, 64 scripted speech, 37 SCSA, see Signal Computing System Architecture (SCSA) SCSA Telephony Application Object, 64 SDT, see signal detection theory (SDT) search, 101, 107, 176, 190, 191, 225, 231, 237, 239, 240, 256, 270-281, 509 greedy, 108 single best sentence, 237, 271-275word graph, see word graph search errors, 271

second language acquisition, 89-90 segmental complexity, 531, 559 segmental quality, 482, 498, 519, 520, 526, 527, 529, 531,534, 541, **542**, 543, **547**- ${\bf 548},\ 548,\ 560$ segmental structure, 83, 152, 533 segmental test, 498, 503, 519-529, 547 - 548at the sentence level, 526-529 at the word level, 520-526 segmentation, 39, 54, 81, 142, 150-155, 172, 206, 346, 363-**365**, 365–367, 382, 402, 407, 512, 517, 530, 584 automatic, 153-154, 318, 363 in Verbmobil, 154-155 manual, 152-153, 153 morphological, 517 of word boundaries, 166 semi-automatic, 153-154 segments, functions of, 519 Semantically Unpredictable Sentences (SUS), 492, 494, 501, 529, 529, 539, 544, 547, 550, 557 semi-automatic segmentation, see segmentation, semiautomatic semi-spontaneous speech, 102-103, 165sensitivity profile, 464 sentence error rate, 390, see error rate, sentence sentence model, 181, 182 sentence syntax, 217-218 sex, see speaker, sex of sex identification, 51, 408, 464 sheep, 43, 394, 432 shimmer, 296, 298, 298 sibilation, 554 sign model, 193-195 Signal Computing Bus, 66 Signal Computing System Architecture (SCSA), 47, 66 signal contamination, 340

signal detection theory (SDT), 367, 368 signal level, 393, 400 signal-to-noise ratio, 19, 45, 48, 50, 70, 98, 122-124, 129, 133, $134,\ 136,\ 294,\ 301,\ 311,$ 312, 327, 384, 390, 391, 393, 399, 424, 478, 500 simple hypothesis testing, 354, 354- $\mathbf{356}$ simple random sample, 348, 349, 362, 363 slips of the tongue, 89 SLM, see Speech Level Meter small room acoustics, 307-309 smoothing, 237, 238, 248-262, 266, 281-287, 499 absolute discounting, see discounting, absolute linear discounting, see dicounting, linear linear interpolation, see linear interpolation SNR, see signal-to-noise ratio SOAP, see Speech Output Assessment Package (SOAP) sociolect, 117, 117, 118 sociolinguistics, 22, 80, 87-89, 90, 99, 105, 111, **116–118**, 146sound level meter, 338, 339 sound pressure level (SPL), 292, 293, 302, 304, 308, 327, 338, 501soundproof booth, 309 SPAR sentences, 362, 362 SPARCstation, 63 sparse data, 153, 188, 191, 238, 241, 253, 262 speaker, 71, 110-118, 394 age of, 22, 36, 51, 71, 79, 87, 111, **113**, 114, 119, 145, $390,\ 394,\ 408,\ 409,\ 413,$ 423, 424, 464, 535, 536, 538applicant, 411-413, 421, 431, 439, 440, 446, 458

cooperative, 415, 417, 421-423, 425dependable, 432, 437, 439discordant, 413 genuine, 411-414, 421, 423, 425, 434, 436, 439, 444-446, 460 mistaken, 412, 414, 464 non-registered, 411 registered, 411-414, 417, 420-425, 427, 428-429, 429-436, 440, 441, 444-446, 450, 451, 458-460, 469 resistant, 433, 437, 441, 448 sex of, 22, 36, 51, 71, 79, 87, 93, 110, 111, **111–113**, 113, 119, 138, 348, 359, 393, 394, 408, 413, 420, 422, 423, 425, 427, 436, 442, 446, 464, 535, 536, 538 stable speaker characteristics, 111 transient speaker characteristics, 111 uncooperative, 422, 423, 425, 477 unreliable, 432, 437, 439, 447 violated, 412, 414, 449 vulnerable, 433, 437, 441, 448 weight/height of, 111, 114, 138. 535speaker adaptation, 36-37, 44, 382, 586speaker alignment, see alignment, speaker speaker change detection, 411, 464, 466 speaker characteristics, 417 speaker class identification, 409 speaker class verification, 409, 410 speaker classification, 408, 408-409, 413, 414, 423 speaker cluster selection, 409, 411 speaker dependency, 35, 35-36, 36, 40, 52, 71, **94–95**, 95, 109, 181, 247, 381, 382, 384, 394, 401, 405, 417, 427,

443, 450-454, 457, 458, $460,\;462,\;466,\;471,\;571$ speaker dependent recognition task, 382speaker identification, 4, 21, 30, 51-**56**, 73, 91, **96**, 409–411, 413, 414, 424, 437, 463, 464, 470, 478 closed-set, 96, 411, 412, 418, 420open-set, 96, 411, 412, 414, 421, 425, 430, **458–459**, 460vs. verification, 410-411speaker independent recognition task, 382labelling, speaker see labelling, speaker speaker matching, 411 speaker misclassification, 412, 412, 414, **431–432**, 434–437, 445, 458-460 speaker population, 436 speaker recognition, 15, 24, 30, 92, **96–98**, 109, **109**, 110, 113, 373-375, 408-410, 417, 418, 432, 463, 465-467, 473, 474, 476-480 speaker recognition system, 97, 98, 410, 412, 413, 418-420, 423, 459, 461, 464 a taxonomy of ~s, 410-418 common-password, 412, 414, 415event-dependent, 412, 414, 420 fixed-vocabulary, 412, 414, 416-417, 418 personal-password, 412, 414, 416, 426 text-dependent, 412, 414, 415-416, 425, 426, 462, 469, 474text-independent, 412 - 415, 417-418, 427, 462, 474 text-prompted, 97, 415, 462, 474

unprompted, 413, 415, 417, 418

unrestricted, 412, 415, 417-418voice-prompted, 413, 415-417 speaker recruitment, 138, 138-140, 141 speaker scheduling, 116, 138, 139, 140 - 141, 141speaker selection, 43, 466 speaker stress, 24, 34, 37, 121, 125, 301, 318, 320, 408, 419, 600 speaker verification, 4, 11, 16, 20, 21, 24, 30, 51-56, 72, 92, 96-**98**, 344, 347, **373–375**, 408 - 480text-dependent, 467, 470 vs. identification, 51, 410-411 speaking mode, 35 speaking rate, 37, 337, 340, 390, 391, **393**, 466, 507, 510, 530, 562speaking style, 18, 83, 85, 99, 117, 129, 165, 290, 337 Spearman's rho, 358 spectral analysis, 110, 112, 124, 339 spectral dynamics, 106 spectral slope, 106, 112 spectral structure, 106 speech activity, 49, 426 speech impairment, 392, 393 speech input/output interfaces, 64, **65**, 75 speech level, 384, 400, 550 Speech Level Meter, 399 Speech Output Assessment Package (SOAP), 7, 496, 506, 550-552, 557, 561 speech processing board, 64, 65, 65, 75speech quality, 111, 327, 383-384, 384, **418-419**, 423, 424, 544speech quantity, 419-420, 421, 424, 425speech recogniser, see speech recognition system

speech recognition, 4, 7, 15, 20, 30, **34–50**, 62, 66–69, 91, 92, 94-96, 96, 97, 113, 166, $178,\ 186,\ 187,\ 197,\ 199,$ 209, 210, 213, 214, 216,218, 224, 231-233, 237, 238, **238-244**, 246, 261, 266, 272, 281, 290, 339, 347, 360, 368, 373, 375, 408, 413, 418, 464, 466, 475, 542, 565, 571, 586, 608 large vocabulary, 237-239, 241, 271, 273, 401small vocabulary, 273 speech recognition system, 7, 16, 20, 21, 24, 30, **34-50**, 63, 66, 70-72, 94-96, 100, 104, 109, 110, 115, 117, 118, 124, 129, 151, 153, 155, 157, 161, 166, 185, 188, 190-191, 203, 204, 206, 212, 216, 224-226, 237-239, 245-247, 270, 271, $299,\ 309,\ 336,\ 340,\ 344,$ 346-348, 360-365, **366**- $373,\ 374,\ 381{-}603$ classification of ~s, 381-383 connected words, see connected words/speech continuous speech, see continuous speech isolated words, see isolated words large vocabulary, 7, 24, 84, 109, 170, 178, 180-182, 197, 239, 270, 271, 381 -384, 390, 395, 396, 401-403, 405, 406 medium vocabulary, 178 small vocabulary, 381, 401, 406 speaker-dependent, 35, 36, 94-**95**, 95, 450 speaker-independent, 35, 36, **94-95**, 95, 109, 181, 382, 384, 394, 401, 405, 427, 450, 466

text-dependent, 54 text-independent, 54-55speech synthesis, 4, 7, 15, 16, 20, 21, 24, 30, **56-61**, 69, 73, 74, 92, **92–93**, 96, 99, 108, $110, \ 151, \ 161, \ 162, \ 165,$ 166, 178, 185-187, 191,197, 205, 210, 213, 217, 218, 224, 233, 291, 295, 339, 344, 359, 360, 373, 375, 377, 392, 464, 466, 481-562, 565, 566, 569-571, 583, 584, 588, 601, 603 speech variety, 420, 424 SPL, see sound pressure level spoken language corpus, 5, 10, 11, 18, 19, 22-23, 23, 24, 79-80, 80, 83, 86-98, 107, 178, 188, 194, 219, 222, 578spoken language data, 18,22vs. written language data, 81-86 spoken language dialogue, 24, 233, 564, 568, 571, 573, 576, 579.581 spoken language identification, 409, 410, 413 spoken language system, 4-6, 11, 16, 16-17, 19-20, 23, 28, 30, 92, 96, 104, 143, 176, 177, 178, 180-183, 189-191, 206, 208, 209, 218, 219, 222, 224, 225, 231, 232, 235, 236, 564, 574, 582small vocabulary, 180 spoken language verification, 409, 410spontaneous speech, 18, 22, 37, 71, 82, 85, 88-91, 96, 99, **103–106**, 109, 113, 118, 121, **125–126**, 126, 127, 144, 146, 147, 147-148, 149, 155, 157, 158, 163, 165, 167, 169, 171,

185, 346, 354-356, 359-361, 383, 384, 413, 415,578, 598 transcription of, 147-148 stability, ${\bf 332}$ standard communication protocols, 67 standard deviation, 350, 352, 356 standard error, 350, 351 standard switching compliance, 47 stem, 183, 186-188, 191, 193, 196-199, 202, 207, 208, 212-215, 215, 216 stem spotting, 213 stochastic grammar, see grammar, $\operatorname{stochastic}$ stochastic knowledge sources, 239 stochastic systems vs. knowledge-based systems, 94 stop, 107, 160, 181, 299 stress, 59, 60, 100, 108, 111, 154, 159, 161, 164, 181, 188, 205, 209-211, 213-215, 218, 220, 227, 299, 486, 490, 499, 501, 508, 509, 514, **515–516**, 516, 517, 519, 525, 530, 531, 534, 545-548, 559 on speaker, see speaker stress structuralism, 573, 574 studio, 22, 48, 120–122, 122-123, 123, 126, 128, 129, 140, 141, 301–303, **309–310**, 310, **310–312**, 312, 314, $\mathbf{316},\ 322,\ 418,\ 480$ stuttering, see pathological speech, stuttering subject age of, 295, 600 sex of, 295, 600 weight/height of, 298 subject variables, see variables, subject sublanguage, 182 substitution, 49, 52, 53, 58, 114, 154, 281, 385-388, 407, 608

subword unit, 39, 40, 386 suffix, 165, 196, 212-214, **215** suprasegmental features, 100, 161, 163SUS, see Semantically Unpredictable Sentences (SUS) sustention, 554 Switchboard corpus, 109, 125, 146, 148, 161, 169-171 syllable, 18, 82, 85, 107, 108, 155-157, 162, 163, 165, 167, 170, 186, 188, 188, 196, 205, 210, **212**, 213, 214, 227, 239, 299, 337, 515, 524, 525, 532, 547 orthographic, 213 stressed, 164, 167, 515, 519, 530, 531, 548 unstressed, 108, 299, 519, 525, 547, 548 syllable boundary, 108, 213 syllable duration, see duration, syllable syllable monitoring, see monitoring, syllable syllable structure, 181, 188, 213, 501, 515, 523, 552, 559 synonymy, 201, 201, 202 syntagmatic distribution, 218 syntagmatic relation, 200, 201, 217 syntax, 34, 40-41, 103, 113, 113, 117, 120, 124, 148-150, 157, 158, 162, 194, 217, **217**, 218, 223, 239, 381, $382, \ 394, \ 395, \ 398, \ 566,$ 586dialogue, 217 finite state, 43 phrasal, 217 sentence, 217-218 word, 214, 217 system capacity, 68 system expertise, 585 system lexicon, 177, 183, 184, 186-187, 231, 234 system model, 598 system simulation, 21, 25, 67, 75

system-in-the-loop, 105, 581 T2 (Europe), 47 talker descriptors, $\mathbf{296}-\mathbf{299}$ talking style, 383, 384, 393 TAPI, 64, 68 task completion rate, 604 task model, 62, 570, 598 task-oriented dialogue, 568 task-oriented speech act, 575 TEI, see Text Encoding Initiative (TEI) template phone, 180 spectral, 180, 181 tempo, 530 temporal drift, **419**, 424 terminal emulator, 397 test corpus, 92, 95-97, 189, 237, 592 test data, 104, 245, 257, 258, 260, **346**, 348, 360, 366, 444, 451, 461, 581 test database, 43, 385, 389-391 test material, 80, 95, 100, 120, 128, 246, 362, 372, 390, 393, 395, 398, 402, 405, **405**-406, 412, 414, 415, 418, 419, 424, 427, 475 test mode, 410, 420, 424, 425 test phase, 142, 410, 415, 420 test procedures, 488, 491, 494-496, 511, 516, 518, 547, 549 test session, 95, 419, 420, 424 test set, 101, 245, 246, 257, 260, 366, 392, 395, 429, 432-436, 438, 440, 442, 443, **443**, 444, 445, 447, 448, 451, 453, 454, 457, 459-461test set perplexity, 246 test utterance, 390, 399, 412, 414, 419, 428-436, 442, 450, 458 - 460test vocabulary, 389-391 testing, 4-6, 24, 87, 95, 96, 98, 101, 109, 110, 124, 151, 204,281, 309, 336, 346, 347,

 $360, \ 361, \ 366, \ 376, \ 389,$ $396, \ \mathbf{398}-\mathbf{401}, \ 410, \ 419,$ 420, 423, 424, 462, 463, 566, 581, 593, 596, 602 automated, see automated testing/assessment black box, see black box comparative, see comparative testing diagnostic, see diagnostic approach field, see field testing functional, see functional testing glass box, see glass box judgment, see judgment testing laboratory, see laboratory testing speech output, 484, 485, 487, 488, 490, 491, 502, 504, 511, 525, 539, 541-550 testing phase, 95-97, 109 testing session, 394, 398 tests χ test, 358 SAM Overall Quality Test, 496, 506, 550, **561** SAM Prosodic Form Test, 531, 559 SAM Prosodic Function Test, 533, **560** SAM Standard Segmental Test, 503, 520, 522, **522**, 523, **523**, 524, 525, 539, 550. 551Bellcore Test, 520, 524, 524, 540, **553** CLuster IDentification (CLID) Test, 520, 523, 524, 524, 525, 529, 539, 544, 547, 550, **552** Diagnostic Pairs Sentence Intelligibility Evaluation Test (DPSIE), **525**, 526 Diagnostic Rhyme Test (DRT), 392, 494, 520, **520**, 521, 521, 522, 525, 547, 554

Diphone Test, 520, 523-524, 524Friedman test, 358Harvard Psychoacoustic Sentences, **527–528**, 528, 529, **558** Haskins Syntactic Sentences, 528, **528–529**, 529, **556** independent t test, 358, 359 ITU-T Overall Quality Test, 507, 508, 536, **562** Jonkheere's trend test, 358 Kendall's W, 358 Kruskaal-Wallis, 358 Mann-Whitney U test, 358, 360 McNemar test, 359, 401 Median Test, 358 Minimal Pairs Intelligibility Test (MPI), 520, 525, 526 Modified Rhyme Test (MRT), 494, 520, **520**, 521, **521**, 522, **522**, 539, 540, 547, 555Page's L test, 358 related t test, 358, 359 Unpredictable Semantically Sentences (SUS), 492. 494, 501, 529, **529**, 539, 544, 547, 550, 557 Spearman's rho, 358 Wilcoxon Sign test, 358, 359 Z test, 358 Text Encoding Initiative (TEI), 148, 168, 169 text preprocessor, 512, 513, 545 text-dependent speaker recognition system, see speaker recognition system, text-dependenttext-independent speaker recognition system, see speaker recognition system, textindependent text-prompted speaker recognition system, see speaker recognition system, text-prompted

text-to-speech (TTS), 4, 56, 58-61, 65, 66, 73, 108, 116, 151, $155,\ 178,\ 187,\ 197,\ 219,$ $226,\ 232,\ 403,\ 481,\ {\bf 481},$ $482, \ \ 485, \ \ 488, \ \ 490, \ \ 495,$ 497, 501, 504, 505, 510, 511, 516-518, 520, 532, 536, 545, 560, 607 time slot, 66 time-frequency warping, 500-502 timing and control signal management, 47tip and ring wires, 47 ToBI, 162, 163, 163-164, 164, 165, 167, 168, 172, 211 tone, 162-164, 166-168, 188, 299, 530tone signalling, 570 total word error rate, 388 touch tone, 46, 568 training, 24, 94, 95, 97, 100, 109, 110, 124, 151, 153, 155, 166, 185, 191, 204, 224-226, 243, 253, 257, 260, 262, 267, 346, 347, 360-363, 366, 382, 383, **383**, 386, 389, 395, 397, 398, **398**, 399, 402, 410, 419, 420, 424, 427, 434, 461, 462, 466, 491, 578, 599 air-traffic control, 565, 599 embedded, 381, 383 of subjects, 492, 495, 519, 521, 528, 529, 532, 537 of wizard, 377, 587, 587, 589 training corpus, 95, 96, 118, 189, 248, 255, 259 training data, 104, 153, 241, 245, 248, 249, 251-253, 255, 257-261, 263, 265, 267, 270, 282, 283, 346, 348, 360, 366, 418, 451, 461, 462, 578, 581 training material, 35, 38, 53, 71, 95, 129, 362, 372, 398, 402, **402–404**, 405, 412, 414, 418, 419, 424, 427, 451,

475training method, 381, 394 training mode, 410, 420 training passes, 383 training phase, 35, 36, 38, 40, 45, 53, $54, \ 95-97, \ 109, \ 239, \ 241,$ $410,\;419,\;420,\;571$ training session, 95, 394, 395, 398, 419, 420, 424, 425, 562 training set, 248, 257, 258, 366 training text, 250, 417 training time, 395, 403 transaction logfiles, 70 transaction success, 25, 610-613, 614transaction success rate, 32, 604, 605 transcription, 4, 17, 22, 23, 26, 36, 39, 54, 59, 80, 81, 83, 84, 89, 91, 117, 120, 129, $146, \ \mathbf{146}, \ \mathbf{147}, \ \mathbf{147}{-}\mathbf{148},$ 150, 152, 153, **155–158**, 158, 159-168, 168-172,188, 190, 194, 199, 203-209, 211, 225, 226, 228, 239, 364, 367, 383, 386, 387, 403, 495, 514, 516, 545, 546, 587, 593 acoustic-phonetic, 150, 160-161 broad phonetic, 84, 150, 159-160levels of, 148, 149 morphophonemic, 194 narrow, 152 narrow phonetic, 83, 84, 150, 152, 160 of dialogues, 148 of read speech, 147-148of spontaneous speech, 147, 147 - 148 $orthographic, \ 54, \ 80, \ 120, \ 146,$ $147, \ \mathbf{147}, \ 149, \ 150, \ 153,$ **155–158**, 158, 161, 170, 171, 183, 190, 191, 204, 225, 226, 234, 383, 403 phonemic, 80, 120, 129, 146, 155, 171, 188, 190, 194,

203, 205, 205, 206, 206, 209, 225, 226, 514, 516, 545 $phonetic, \ 36, \ 39, \ 59, \ 80, \ 84,$ 120, 146, 147, 170, 171, 194, 209, 239 phonological, 204, 207 phonotypic, 150, 159-160, 209 physical, 150, 161 prosodic, 23, 150, 161-168, 172types of, 148, 149 transduced signals, 80, 119 transducer, 289, 301-321, 336, 338 transducer characteristics, 45 transliteration, 117, 129, 142, 146, 191transmission factor, 333, 333 trash model, 42 treatment level, 357, 357 trie, 217, 221 trigram, 43, 94, 187, 237, 238, 241, 243, 244, 246, 248, 249, 251, **253–262**, 264–270, 280, 403, 404 phone, 243 triphone, 60, 102, 181, 182, 239, 243, 402, 403 TSAPI, 64, 68 turn, 96, 125, 126, 134, 148, 170, 378, 379, 564, 568, **568**, 573, 576, 577, 592, 596, 598, 608-610, 614 turn duration, 605, 606, 614 tying, 255, 255, 256, 262, 286 type hierarchy, 192 ultradirectional microphone, see microphone, ultradirectional ultrasound imaging, 132uncooperative speaker, see speaker, uncooperative unidirectional microphone, see microphone, unidirectional unigram, 43, 241, 251, 252, 254, 257-259, 261, 395, 404 uniqueness point, 179

unprompted speaker recognition system, see speaker recognition system, unprompted unreliable speaker, see speaker, unreliable unrestricted speaker recognition system, see speaker recognition system, unrestricted unseen events, 243, 248-253, 254 untraining, 410 untraining mode, 410user expectations, 32 user model, 598 user profile, 62 user-belief module, 62 usurpation, 422 usurper, 413 variable, 348 dependent, 348, 375 independent, 348, 357 variables communication channel, 588-589 dialogue model, 587 in spoken WOZ experiments, 583 - 589subject, 377, 583-585 knowledge variables, 584 production variables, 584 recognition variables, 583 Wizard, 377-378, 586-588 production variables, 586 recognition variables, 586 verification, see spoken language verification vernacular, 87, 105 violated speaker, see speaker, violated violation, 422, 423 ${\bf visual \ \, environment}, \ \ see \ \ environ$ ment, visual VME-based system, 63, 470 vocabulary, 34, 38-41, 41, 42, 44, 71, 103, 113, 113, 117, $118,\ 120,\ 121,\ 124,\ 127,$ 129, 181, 182, 185, 190,

214, 219, 239-241, 243, 245-247, 256, 257, 266, $273, \ \ 280, \ \ 377, \ \ 381\text{--}383,$ $383,\ 384,\ 385,\ 387,\ 388,$ **388–389**, 390, 393, 395, 402, 405, 470, 487, 503, 566, 570, 579, 580, 584, see asolexicon 857application, 39, 42, 42, 71 closed, 197, 246 fixed, 39, 40, 412, 414, 416-417, 418 flexible, 39, 40 large, 7, 24, 84, 109, 170, 178, 180-182, 185, 197, 198, 213, 216, 237-239, 241, $270,\ 271,\ 273,\ 281,\ 381-$ 384, 390, 394-396, **401**-406 medium size, 199, 381 open, 551, 552 size of, 34, 95, 198, 203, 209, 215, 221, 241, 247, 248, 381, **381**, 393, 395, 403, 566, see asolexicon size857 small, 24, 38, 180, 197, 273, 381, 392, **396-401**, 402, 591structure of, 185 unlimited, 215, 247, 381, 394 wizard, 584 vocabulary coverage, 246, 251 Vocal Profile Analysis Protocol, 537 voice characteristics, 97, 336, 466, 468, 469, 479, 482, 500, 503, 518, **535-538**, **549-**550functions of, **535–536** voice characteristics tests, 536-538 voice masking, 422, 464 voice personality, 536 voice stop, 48, 49 voice training, 22, 116, 296, 360 voice-prompted speaker recognition system, see speaker recognition system, voice-prompted

voicing, 107, 115, 147, 153, 160, 297, 363, 415, 492, 493, 554 vulnerability, 444-445vulnerable speaker, see speaker, vul- $\mathbf{nerable}$ Wall Street Journal, 96, 102, 237, 257, 258, 383, 403-405 waveform coding, 294 weight, see speaker, weight/height of white noise, 292, 325, 334, 337, 500 whole word model, 39, 40 Wilcoxon Sign test, 358, 359 within groups variance, 357 Wizard of Oz (WOZ), 18, 25, 67, $104, \ 104{-}105, \ 143, \ 145,$ 375, **375–378**, 475, 569, 581 - 591Wizard variables, see variables, Wiz- ard wolf, 441 word graphemic, 182 morphological, 196, 197 orthographic, 163, 164, 196, 516phonetic, 182 phonological, 162, 196 prosodic, 196 syntactic, 197 word boundary, 84, 166, 167, 205, 272, 275-280, 382, 402 word classes, 191, 218, 237, 262-265, 270, 574 word clusters, 265 word error rate, 390, see error rate, word word formation, 214 word graph, 190, 203, 237, 275-**281**, 281 word label, 44 word lattice, 190, 191, 203, 212, 237, 277word model, 181, 182, 209, 216, 239, 279, 383, 393 stochastic, 177, 185, 225

word recognition, 20, 95, 177, 182, $187,\ 213,\ 239,\ 272,\ 382,$ $386{-}388, \ \ 394, \ \ \mathbf{396}{-}\mathbf{401},$ $401,\ 402,\ 407,\ 519,\ 530,$ 534, 548, 596 word spotting, 40, 44, 95, 178, 387, 392, 570, 598 Write Once Read Many (WORM), 136written language data vs. spoken language data, 81-86 WSJ, see Wall Street Journal X-ray microbeam, 132 z score, 351, 353 Z test, 358zerogram, 241, 254ZIP-codes, 102