German Project

Source:

Dafydd Gibbon
Tel: +49-521-106-3510/09
Email: gibbon@asl.uni-bielefeld.de
Fax: +49-521-106-5844

Gunter Braun
Michaela Johanntokrax
Michael Schwalbe
Tel: +49-521-106-5274

Universität Bielefeld, Fakultät für Linguistik und Literaturwissenschaft
P 8640, D-4800 Bielefeld 1

Contents

Abstract

This report consists of two parts:

A short presentation of a schedule - based on linguistic and computational linguistic techniques - for speech database design (prediction of contents) and speech data description. Different levels of transcriptions with defined symbol sets will be exemplified using German data.

The production of the actual German speech database - the recording environment, the text corpus, speaker characteristics, and recording procedure - is described next.

The aims of the report are to enable speech researchers and phoneticians to make full use of the German part of the multi lingual speech database EUROM.1.

2 Database Design

The production of speech databases, that is the recording and annotation of speech data, is time consuming and expensive. In order to make maximal use of existing resources it is highly important to be able to predict speech database contents at various levels before recording and thus design specific task oriented speech databases.

In case of multi-lingual speech databases compatibility between different language corpora is only ensured if their predictions are comparable. Predictions can be made on various levels:

- quality of recording (depending on recording conditions)
- speaker characteristics (age, sex, etc)
- linguistic dimensions

For the production of the multi-lingual speech database EUROM-1, a recording protocol was prepared in order to enable the production of standard recordings and speaker selections.

The text corpora for the individual languages consist of a fixed number set, comparable logatomes, and cognitively linked short passages. The short passages have been translated from an English text corpus to resemble each other in syntactic and semantic complexity for different languages.

For phonetic predictions of orthographic (prompting) texts, phonological transcriptions can be used. Phonological transcriptions, which are enclosed in slashes //, reflect not only individual sounds but also the sound system of a given language and thus enables an abstract description of different pronunciation variants. A

computer readable phonemic symbol set (SAMPA, see APPENDIX I for German SAMPA) has been defined as an ASCII encoding of an IPA subset for 8 European languages (Wells, 1989).

The final sound prediction - the phonotypical transcription of a text in a given language (enclosed in []) - describes the standard pronunciation, where all standard word level assimilation rules have been applied and phonetic variants (allophones) which have no corresponding phonemic symbol may be included.

Phonotypical transcriptions can be analysed (using the software tool SAMTRA developed within SAM) to check distributions of predicted sounds resulting from different orthographic texts to ensure enough occurrences of each individual sounds or sound combinations in a speech database.

The production of a German phonotypical transcription is shown by the following simple example. The orthographic text, which might be the prompting text in an anechoic room:

In Burgen rasseln Gespenster oft mit Ketten.

would have the phonological representation (were only phonemes occur):

/ In bURg@n Ras@ln g@SpEnst@R Oft mIt kEt@n /

To predict the standard pronunciation in German, the phonemic symbol set has to be extended to cope with allophonic variants (see appendix). The phonotypical transcription:

[In bU6gN Ras@ln g@SpEnst6 Oft mIt kEtn]

reflects standard assimilations.

In /bURg@n/ the consonant /R/ in the vowel consonant combination: /UR/ is vocalized to [6] and constitutes a diphthong [U6] with the preceding vowel /U/; the elision of the schwa vowel /@/ causes the velar assimilation of /n/ to [N];

In /g@SpEnst@R/ the consonant /R/ in the schwa vowel consonant combination: /@R/ is vocalized to [6]. The single allophone [6] mostly replaces [@6];

In /kEt@n/ the elision of schwa vowel /@/ leads to [kEtn].

Assimilations and elisions across word boundaries are hard to predict, when texts are carefully read aloud in an extreme recording situation; therefore they are not used in German phonotypical transcriptions.

Fig.1 sketches the described sound predictions and database design in the context of speech database production tools within SAM standards.
On the left hand side, required resources:

- linguistic knowledge for database corpus design,
- production of speech in the recording situation, and
- phonetic experts (or semi automatic labelling systems using phonetic knowledge) for annotation purposes are given, whereas SAM database files are listed on the right hand side. The process and dataflow is represented by arrows.

3. Database Production

3.1 Recording Environment

The recordings took place in an anechoic room of about 9 square metres. As shown in fig.2, the recording room contained a table on which a VGA black and white prompting monitor was placed slightly angled in regard to the speaker; a chair for the test person, a small loudspeaker for communication, and the B&K 4155 microphone. In case of the Few Talker and Very Few Talker recordings, a laryngograph from Laryngograph Ltd. was added to the equipment.

The lighting was set up in such a way that there were no distracting light reflections on the prompting screen. To minimize stress factors for the speakers, the chair had no device for head fixation. The microphone was placed at a distance of about 50 centimetres to the speaker, and there were no reflecting surfaces or sound absorbing materials between the test person and the microphone.

The data connection between operators room and the anechoic room consisted of the following cables:

- The B&K extension cable AO 0028 (connection between the defined B&K sound level meter 2235 and the microphone 4155).
- A VGA cable (connection between the SESAM workstation and the VGA monitor).
- A two core cable (connection between the amplifier and the loud speaker).
- A coax cable (connection between the DAT and the laryngograph).

3.2 Database Text Corpus

The prompting files for the German Eurom-1 recordings consist of three types which reflect important topics in speech recognition: number recognition, single word recognition and continuous speech recognition.

(i) Numbers:

100 numbers in the range of 0 to 9999 grouped to 5 blocks of 20 numbers (see APPENDIX II).

Subjects received no instructions regarding the pronunciation of numbers, except for those which could be read as dates, e.g. 1919 should be pronounced "eintausendneunhundertneunzehn" not "neunzehnhundertneunzehn". The numbers were presented as a string of digits and not orthographically to minimize stress and error rates, as subjects had difficulties with reading numbers written out as words in a limited time, especially as numbers regardless of their length form single orthographic words in German, e.g. "dreitausendsiebenhundertsiebenunddreißig".

(ii) CVC Words:

72 (C)CVC words grouped to three blocks of 20 and one block of 12 (see APPENDIX III). The CVC words were read additionally by the Very Few Talker Set in a context of two words to get well defined data for coarticulation research. The CVC word surrounded by the context words forms a simple German sentence. The syntactic structure is VERBimperative NOUNname ADVERB. Reading sentences minimize specific prosodic phenomena which occur when reading single word lists. A fixed syntactic structure holds for a comparable prosodic contour.

(iii) Passages:

40 texts consisting of 5 sentences (see APPENDIX IV). Each Passage was translated from the English original (produced at UCL). Modifications were made to ensure a defined distribution of German sounds using a cyclic scheme of segmental content predictions and reformulations of the passages. The final passages provide sufficient sound distributions - additional sentences are not needed.

For all prompting files phonotypical transcriptions were made for segmenta prediction purposes. A simple lexicon consisting of orthographic, morphologic, phonological, and phonotypical entries was used to ensure identical phonotypical representations for multiple occurrences.

The orthographic and phonotypical entries of the lexica for the numbers and the CVC words can be found in Appendix II and Appendix III respectively.

The orthographic passages with their corresponding phonotypical transcriptions are listed in APPENDIX IV.

All phonotypical transcriptions were verified and analysed using the software package SAMTRA. The resulting sound predictions of all passages can be found in Appendix IV.

3.3 Speaker Selection

Following the Recording Protocols, the speakers should be as varied as possible, and a wide range of accents should be obtained in the database. The speakers for the German EUROM.1 recordings have been chosen from university staff and their social contacts. Therefore, they mostly have similar social backgrounds. The actual database contains 30 female and 33 male speakers; some phonetically relevant personal data (age, sex, weight, height, smoking habit) are listed in the EUROPEC standard description file SPEAKERS.DBF.

The distribution of the German test persons in regard to their age and sex is given infig.3; speakers from the Few Talker Set are underlined, speakers from the Very Few Talker Set are marked by double underlining.

A total of 63 test persons were selected for the database recordings (three more than requested, as to have the ability to replace defective speech files if necessary).

For the Many Talker Set recordings, all 63 speakers had to read 100 numbers and 5 randomly chosen but regularly distributed passages (see Appendix VI). From this set of speakers, twelve were selected to perform the Few Talker Set recordings. They had to read out the whole set of 100 numbers and the CVCs five times plus 20 passages in blocks of four each. Two of these twelve speakers formed the Very Few Talker Set. They additionally read the CVC words within the context words once and the context words themselves five times.

3.4 Recording Procedure

It was decided that recording sessions should be supervised by two operators in order to avoid fatigue and to ensure a sufficient control over speaking errors.

Each recording session started with a informal talk with the speaker to prevent stress and create a pleasant atmosphere. Each speaker got a brief description of the project and its aims. The speakers were invited to have a look at the anechoic room as long as they wanted to get used to its unusual acoustic characteristics. The possibility of stopping the recording session at any time by metacommunication with the operators was offered.

The speakers were informed about the duration of the recording session, how to behave during sessions (e.g. to keep a fixed position as far as possible), and they were told not to take watches into the anechoic room. The speaker's personal data

were taken and written into the SPEAKERS.DBF while it was promised to delete their names in case of distribution (in accordance with German law). A standardized instruction about the prompting style on the monitor and the number and forms of the promptings was given next. One operator guided the speaker into the anechoic room, checked the appropriate distance of the microphone, and in case of the Few Talker and Very Few Talker Sets he helped the speaker to place the laryngograph neckband correctly.

As a 60 MB hard disk was used, the recordings of the Few Talker Set had to be split into five sessions of about 20 minutes. To backup the data on a streamer tape, a pause of 25 minutes was necessary after each session. This resulted in a total time of three and a half hours for each Few Talker. All Few Talkers preferred their recording to be made on one morning or afternoon. The additional time amount for the Very Few Talker Set was about half an hour. In case of the Many Talker Set it was possible to store four speakers' data on the hard disk. Therefore the backup break had to be made after one hour recording time (15 minutes for each speaker session).

After each session about 5% of the speech signal and the corresponding orthographic label files was checked for quality.

Additionally, all Many Talker files were checked entirely before the final backup to Exabyte.

References

Autesserre, D., Perennou, G., Rossi, M. (1989): Methodology for the Transcription and Labeling of a Speech Corpus, in: Journal of the International Phonetic Association, 19:1, pp. 2-15.

Braun, Gunter (1991): SAMTRA (VS 1.0) SAM Transcription Analysis. Program documentation. UBI-SAM-3/91.

Gibbon, Dafydd (1991): Linguistic Aspects of Speech Material Complexity. UBI-SAM-1/91.

Kohler, Klaus (1990): German, in: Journal of the International Phonetic Association, 20:1, pp. 48-50.

Tomlinson, M.J. (1990): Guide to Database Generation - Recording Protocol. SAM-RSRE-15 University College London (1989): Levels of Transcription. SAM-UCL-Draft.

Wells, J.C. (1989): Computer-coded Phonemic Notation of Individual Languages of the European Community, in: Journal of the International Phonetic Association, 19:1,

pp. 31-54.