Standards for spoken language

Dafydd Gibbon (draft 05, 2015-03-11)

The terms ‘spoken language’ and ‘speech’ characterise domains of research and application in several disciplines, from phonetics, language teaching and documentary field linguistics through sociology, psychology and speech pathology, some of which have become associated with the meta-discipline of digital humanities, to computational models of components of spoken language and speech technology, each with their subdisciplines, and each with their theories, models, terminologies and *de facto* or institutional standards for best practices. The problems which arise from this multidisciplinary diversity are considerable: the institutional standard ISO 639-3 codes for the identification of languages are a starting point for shared information, but are still rarely applied, even publications in linguistics, phonetics and the speech technologies. There are few institutional standards, when the spoken language domain is seen as a whole, and the field is largely in flux, but there are *de facto* standards and trends.

The present outline of standards for spoken language will first characterize the domains and properties of spoken language as a basis for further discussion, then outline standards developments for basic resources shared by a number of disciplines, such as transcription and speech signal annotation, and for the development and quality control of spoken language resources. The speech technologies of automatic speech recognition, text-to-speech synthesis, and language and speaker identification are not treated in detail here; rather, the focus is more on linguistic and phonetic requirements and standards which are relevant to computational scenarios.

The domains of spoken language: standards versus diversity

Spoken language domains cover a wide range of communication styles, genres and scenarios: communication styles (from intimate through informal to formal), genres (e.g. interview, joke, narrative, public speech, sermon) and scenarios (monologue, face-to-face, audio and video phone, one-way mass media). Historically and in child language development, speech precedes written language, and may itself be predated by gestural communication (McNeill 2000; Gibbon 2011; Rossini 2012). Indeed, speech is a form of gestural communication transduced into the acoustic medium, just as writing, at the physical level of manuscript production, is a transduction of gesture into visible inscriptions. Each modality has different consequences for communication speed, support by memory and cognitive processes, distance coverage in space and durability in time.

The speech-text modality differences also have practical, scientific, ethical and forensic consequences (Gibbon et al. 1997; Gibbon et al. 2000; Austin 2007). Speakers, unlike writers, are often instantly recognisable within fractions of a second, yet their speech is not durable unless recorded on a technical medium. In many scenarios speech is temporally and locally coextensive with gestural and tactile communication modalities; in other scenarios the modalities are separated (e.g. in visually or acoustically challenging situations), or the speech setting is subject to dislocation in speech at a distance (teleglossia, e.g. in telephony and visual conferencing) and distemporality (e.g. in writing). Speech is increasingly seen as multimodal, together with gestural and tactile interaction, and multimodal speech in technical communication has become a major subdomain (Mehler et al. 2012).

The speech-only communication domain is typically found in the oral societies which remain in some parts of Africa, South America and South East Asia, studied by field linguists, ethnologists and anthropologists, often in cooperation with other disciplines such as musicology (Austin 2007). A large part of daily communication in industrially and economically developed societies is substantially similar, though complemented by complex varieties of communication in technically transmitted media, from writing, whether with pencil, stylus, phone or PC, or multimodal internet telephony. Influential scientific conference series such as *Interspeech* (mainly speech engineering), the *International Congress*

of Phonetic Sciences (mainly the physical modality aspects of spoken language) and the *Language Resources and Evaluation Conference* (*LREC*) bear witness to the diversity not only of the domain but of methodologies, and many conferences and journals in other disciplines give implicit or explicit coverage to spoken language.

There is thus no single spoken language research, development and application community, as the present discussion shows, and consequently *de facto* standards for data, tools and information interchange have developed differently in the different communities, and sometimes even basics like phonetic transcription are not uniformly practised. Another factor which militates against the development of comprehensive sets of standards is the complexity of the field and the disparity of topics and R&D interests:

1. Spoken and written language differ not only in the phonetic and prosodic modalities and their levels of representation, but also in the lexicon (e.g. levels of style; hesitation phenomena and other discourse particles), the grammar (e.g. levels of style, rarity of centre-embedding except in formal styles, disfluency handling strategies), and at discourse levels (e.g. turn-taking, turn overlap).
2. In crucial respects the semantics and pragmatics of spoken and written language differ (e.g. in deictic and utterance act properties).
3. Spoken language occurs concurrently and coordinated with visible gestural and postural communication (for a recent account, cf. Rossini 2012) and is itself gesture.
4. Quality criteria, size, accessibility, ethical and legal status of spoken and written data differ.
5. The tools for processing spoken language at the phonetic levels (production, transmission, reception) are specialised and only comparable with the tools for studying written language in terms of manual gesture in handwriting production, typing and touchscreen input, and with the optical character and layout recognition of handwritten and electronically formatted manuscripts and touchscreen gesture signals.

In spite of the speech-text differences, lexical properties of spoken language can in general be catered for by existing lexicographic conventions, and grammatical properties by existing tagset and Treebank conventions, except for the lattices used to represent word hypotheses in speech recognition or turn overlap in discourse analysis. For an ISO standard for dialogue act categories cf. Bunt et al. (2012).

Spoken language resources: transcription standards

Spoken language has specific characteristics at all ranks of linguistic description from speech sounds through phonemes, morphemes, words, phrases and sentences, to utterances and discourse. Compared with constituents of text, units at each of these ranks have their own properties of interpretation, both semantic and phonetic. Semantic interpretation ranges from bare contrastivity of phonemes, through morphemes and words as predicates and operators, to sentences as propositions, texts as argumentation and discourse as negotiation. Phonetic interpretation ranges from sequential segmental consonantal and vocalic patterns and their hierarchical organisation in syllables and larger groups to concurrent prosodic (suprasegmental) rhythmic and melodic features such as phonemic tone, morphemic tone, accentuation, and higher ranking intonational and rhythmic patterning at sentence and discourse levels.

While there are institutional standards for *transliteration* (i.e. the conversion of one system of writing into another, e.g. ISO 9 for Cyrillic or ISO 15919 for Indic scripts), there is currently no ISO standard for phonetic and phonemic transcription. However, professional curating of standardisation in the phonetic and phonemic representation of language is administered by the International Phonetic Association, and the alphabet, including diacritics, has a complete Unicode encoding.

There is one outstanding set of professional *de facto* standards which is used in all of the spoken language communities, from linguistic theory and fieldwork research to applications in language

teaching and speech pathology to the spoken language technologies: the IPA[1] , the IPA character coding according to the Unicode standard, and the formulation of descriptive rules for phonetic processes, such as assimilation, based on the IPA. The IPA is an empirical standard, and has evolved as empirical knowledge has developed, with extensions for specialised purposes such as speech pathology. The IPA was originally conceived as an alphabet which can represent all speech sounds which are contrastively phonemic in all languages of the world. The current understanding of the IPA is more phonetic, and the alphabet is intended to represent all identifiable speech sounds, whether contrastive or not. For the representation of phonemes in languages with less common IPA characters, very often these are substituted with no loss of information (if properly defined) by more common characters which are easier to type.

The *International Phonetic Alphabet* (*IPA*) has been curated since 1886 by the main professional body in phonetics, the *International Phonetic Association*[2] (also *IPA*). The segmental categories, characters and glyph sets of the IPA are widely accepted as a standard point of reference, but there are many specific application oriented variant alphabets. Divergent segmental transcription conventions are used in the historical philologies and in anthropological language studies. Extensions of the IPA have been proposed for specialised use cases, for example in speech pathology (Teoh & Chin 2009).

Although the IPA is fully specified in Unicode, IPA codes are scattered over a number of code blocks, presumably for the sake of space economy, where particular symbols are used in the official orthographies of various languages (e.g. 'θ' in the Greek block, or 'ð' in the Latin-1 blocks). This dispersion of characters frequently leads to uncertainty and inconsistency in use by picking similar but differently coded characters. The lack of a coherent use case semantics for code block allocations in Unicode in order to overcome this dispersion property has received some criticism (Hughes et al. 2006). Although many fonts now implement the IPA Unicode characters, many still do not or are proprietary. For this reason, in linguistics the Gentium[3] font of the SIL is frequently used and often recommended for publications.

In the speech technologies a number of keyboard friendly encodings of the IPA have been developed, the most widely used being the *SAMPA/X-SAMPA* (*Speech Assessment Methodologies Phonetic Alphabet*, the 'X' stands for 'eXtended'; cf. Gibbon et al. 2000). The SAMPA/X-SAMPA coding was originally developed in a EU project as an international consensus of speech engineers and phoneticians for easy information interchange. The SAMPA/X-SAMPA alphabet, being a one-to-one encoding of the IPA, is widely (though not exclusively) used internally in system development in preference to Unicode (ISO 10646) for practical reasons, mainly for being human readable and keyboard friendly and not requiring UTF-*n* codecs. Another reason is that Unicode development focuses on rendering on print output devices rather than on efficiency in character input, and print is not always relevant in spoken language computing contexts.

Symbol sets for prosodic transcription are characterised by much greater diversity, which starts at the level of phonemic tone, with numbers 1 to 5 for Mandarin tones, through the accent diacritics ´, `, ^ in Africanist linguistic usage for high, low, high-low etc. tones (and the same diacritics for rising, falling, rising-falling, etc. in intonational pitch contours), to the IPA symbols for tones. These prosodic transcription notations represent categories. In experimental phonetics and speech technology, a categorial system, *ToBI* (*Tones and Break Indices*) has become widely used, though it has limitations for tone languages on the one hand and discourse intonation contours on the other. A relational transcription, e.g. *IntSint* (mainly applied to speech synthesis; Hirst & di Cristo 1998), which represents pitch ranges and pitch changes within a coherent acoustic model, has a different semantics in the phonetic domain from the categorial systems. There are many other systems of prosody transcription

[1] https://www.internationalphoneticassociation.org/content/ipa-chart
[2] https://www.internationalphoneticassociation.org/
[3] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=gentium

besides these, some of which are based on explicit models of speech production or perception, which will chiefly interest specialists in phonetics, psychoacoustics and speech technology.

As with many standards, there are limitations on practical use cases for the IPA. The IPA standard is particularly relevant for the display of IPA characters on screen or printed page. Although IPA is easy to write by hand, there is currently no accepted standard for keyboard input. The main methods are:

1. *ad hoc* keyboard short-cut tools for IPA subsets,
2. internet character selection tables, conversion tools and online keyboards,
3. menu based character tables in word processors.

Perhaps the most ergonomic method for manual input to use the SAMPA keyboard-friendly encoding, and to copy and paste using a converter from SAMPA/X-SAMPA (ASCII) to IPA (Unicode). Tools for all of these methods are easily found on the internet; no specific addresses are given since fluctuation is high. An optimal solution would be a touchscreen display based on the standard IPA chart, either on-screen or as an an 'IPA mouse'. Currently there is much discussion on these unresolved issues and the challenge remains open.

In computational linguistic and software development environments, the internal representation of IPA characters as Unicode or SAMPA or in other internal codings is not an issue as any of these can be easily handled with a conversion table, as the encodings are biunique; the issues are concerned with user interfaces. Very common in a number of technological contexts are also lexicon and rule-based grapheme-phoneme converters for specific languages. The current standard format for text data storage, including IPA, is to use XML with Unicode entities, as in other domains, and the integration of spoken language information into XML formats on this basis is unproblematic.

Spoken language resources: annotation standards

The structural and functional markup notations of Natural Language Processing, such as part of speech or dialogue act tagging (Bunt et al. 2012) are frequently referred to as 'annotation'. The term 'annotation' has a somewhat different meaning in the spoken language technologies and in empirical studies of spoken discourse, where it refers to the assignment of time-stamps aligned with the speech signal to transcription symbols or to structural and functional markup.

Before annotation types which also apply to writing (part of speech tagging, tree structure annotation, etc.) are applied in the spoken language domain, modality specific annotation is required. The speech signal is recorded digitally and annotated manually, semi-automatically or automatically using appropriate tools, by assigning transcription labels to time-stamps aligned with the signal. A distinction is commonly made between *segmentation*, i.e. the assignment of boundary time-stamps to speech signals, and *labelling* or *annotation*, i.e. the assignment of a transcription symbol to interval or point time-stamps. The distinction is parallel to the traditional 'segmentation and classification' procedures in linguistic data treatment. There are currently no general institutional standards for speech signal annotation, but a number of widely used *de facto* standards for specific purposes have emerged.

Formal definitions for annotation systems were given by Bird & Klein (1990) and applied to annotation by Gibbon (2006). More general *annotation graphs*, applicable to both text and speech markup, were defined formally by Bird & Liberman (2001). Summarising: A spoken language annotation A has two hierarchical levels:

1. A set of information tiers (vectors, streams) T of labels $L_1, \ldots, L_n$, each T representing different information about the speech signal (e.g. phonetic information such as speech sounds, tones, intonation, syllables, words, structural information such as parts of speech or functional information such as discourse functions).
2. Each label L is a pair $<E, S>$ of an event representation E, i.e. a transcription symbol, and a time-stamp S, which is a representation of either an interval I or a point P. The interval I may

be understood either as a pair of start and end points P_s and P_e, or by a point P_s and a duration D, or a duration D and a point P_e. The point representations are timestamps.

The implementation of annotation data types varies considerably. An early data type was dyadic, a *pair* of a transcription symbol for an event, paired with a single time-stamp for the interval start (and often system-specific codes, e.g. for colour representation in screen visualisations). A constraint on this pair annotation data type is that the speech recording must, in principle, be exhaustively annotated, otherwise interval ends are unspecified. A different dyadic data type is point event and time-stamp, which has a different temporal semantics from the symbol plus interval start time-stamp.

The most common speech annotation implementation is a *triple* consisting of a transcription symbol and two time-stamps, for the start and end of an interval). The triple annotation type permits partial annotation of a speech signal, since each annotation interval is fully specified. A specialised type of triple system is used for diphone-based speech synthesis, where the semantics of the event is different from other systems: the 'event' is defined as extending from the temporal centre or acoustically salient peak of one speech sound to the centre or peak of the next. A variant which has been used in speech synthesis has a quadruple format: the label, and three time-stamps for start, centre or peak, and end of the interval.

There are two main use cases for spoken language annotation: first, in speech technology, where annotation is primarily fully automatized and based on machine-learning principles; second, in linguistic phonetics and linguistics from phonology to discourse analysis, where annotation is typically manual, using annotation visualisation tools, and annotation mining for descriptive purposes is semi-manual and often spreadsheet based. The following discussion will concentrate on the linguistic use case. There are several high quality and widely used tools which are available for phonetic annotation, some for transcription alone (e.g. *Transcriber*), some in a phonetic workbench (e.g. *Praat*, *Wavesurfer*, *Annotation Pro*), and others in a multimodal annotation environment (e.g. *Elan*, *Anvil*).

The *de facto* standard annotation tool for linguists and linguistically oriented phoneticians is the *Praat*[4] phonetic workbench (Boersma & Weenink 2001), though new annotation tools with enhanced analysis facilities are continually appearing. New developments in providing automatic annotation for linguistic purposes are also appearing, and will lead to the development of new and more efficient workflow practices in this area (e.g. SPPAS, Bigi & Hirst 2013).

Non-computationally interested users are usually interested in the visualisations provided by the tools, not the internal and interchange formats used by these tools, and in the manual or automatic methods for deriving linguistic and phonetic descriptions from the annotations.

Currently the most common formats for information interchange of manual annotations in computational contexts are textual, with either character separated value (CSV) format of an annotation triple <label, timestamp, timestamp>, or the 'TextGrid' format developed for the Praat phonetic workbench (Boersma & Weenink 2001), both dating from pre-XML days. For timestamps, the Praat format uses seconds in a decimal format, while some other formats use milliseconds. The CSV formats can be enhanced *ad hoc* by a metadata header using comment lines. The Praat format has been criticised for not including provision for extensive metadata. The Praat format has each information item on a separate line, and may be represented in a generalised form by the following expression (without regard for line formatting):

$$\texttt{metadata tiercount}_n\ \texttt{(tiername intervalcount}_m\ \texttt{(timestamp}_i\ \texttt{timestamp}_{i+1}\ \texttt{label)}^m\texttt{)}^n$$

The expression is not strictly a regular expression because of the dependency between the subscript and superscript *n* and the subscript and superscript *m*, and the temporal immediate precedence constraint between the subscripts *i* and *i*+1. The definition also applies, at this level of generality, to the main features of CSV formats.

[4] http://www.praat.org/

So far there is no agreed XML standard for speech annotation, though several tools provide export into XML formats. For general computing and archiving purposes, standard CSV formats with metadata comments, and column and row headers are at least as perspicuous as the more verbose formats.

For conversion between formats and for speech annotation mining and manipulation many tools are available (e.g. the online Time Group Analyzer[5] (TGA) Jue & Gibbon 2013), Python modules (e.g. TextGrid Tools[6], Buschmeier & Włodarczak 2013), and many Praat scripting applications[7].

Outlook: technology, quality assessment and standards convergence

The major venues for the dissemination of results in standards development for spoken language systems are the series *Interspeech* and *LREC* (*Language Resource and Evaluation Conference*), while the *COCOSDA* (*International Coordinating Committee for Speech Databases and Assessment*) initiative, particularly the annual conferences of its East Asian Branch, *Oriental COCOSDA*, plays a role in focussing attention on standards for resources and system development in the speech technologies.

For practical purposes, different speech technologies may be distinguished, for which different standardisation requirements are needed, the main technologies being automatic speech recognition (ASR), text-to-speech synthesis (TTS), language identification and speaker identification. There are several ISO and national standards which refer to quality control aspects of these systems, particularly in safety relevant environments, such as the audibility of announcements in acoustically challenging scenarios such as underground train stations and on speech in telecommunications transmission systems, such as GSM encoding, and other acoustic encodings such as WAV, WMA and MP3. Reference may be made to the standard handbooks for information on relevant standards for technical communication (e.g. Gibbon et al. 1997, Gibbon et al. 2000, Mehler et al. 2012).

Although the current situation in the field of spoken language resources, in particular databases and tools, is very heterogeneous, there are nevertheless factors which are gradually leading to convergence in the interests of resource quality and information interchange, the main pressures predictably being the need for reusability of data and the interoperability of tools.

There several national and international centres concerned with the assessment of the quality of speech databases, mainly in the context of data exchange for speech technology research and development (e.g. ELRA/ELDA, Paris), and there is a great deal of ongoing work on inter-transcriber and inter-annotator reliability and consistency. The work on consistency parallels, to a large extent, work on text markup reliability and consistency assessment, except that annotation also has the property of being time-aligned, so that variations in the centisecond region need to be assessed as similar or dissimilar. The studies by Breen et al. (2012) and Szymański and J Bachan (2012) of inter-annotator agreement for two prosodic annotation systems demonstrate current evaluation methodss.

The second major influence on convergence towards shared standards is the use of *de facto* standard interoperable software tools whose formats and visualization provide benchmarks for the development of future resources.

There are signs in current internet discussion, conference contributions and institutional standardization initiatives that collaboratively motivated standards for spoken language are emerging in the following areas:

1. Transcription: IPA, in spite of small divergence for specific application areas, as a durable transcription standard.
2. *De facto* 'favourite' standards for annotation tools and formats, e.g. Praat, though new tools for other use cases and with more facilities are continually emerging.

[5] http://wwwhomes.uni-bielefeld.de/gibbon/TGA/
[6] https://github.com/hbuschme/TextGridTools/
[7] http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html

3. Standards for spoken language database quality assessment in terms of comparison algorithms for different domains.

References[8]

Austin, P. K. and L. Grenoble (2007). Current Trends in Language Documentation. *Language Documentation and Description* 4. London: SOAS, 12-25.

Bigi, B. and D.J. Hirst (2012). SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. *Proceedings of Speech Prosody*. Shanghai: Tongji University Press.

Bird, S. and E. Klein (1990). Phonological Events. *Journal of Linguistics* 26:33-56.

Bird, S. and M. Liberman (2001). A formal framework for linguistic annotation. In: *Speech Communication - Special issue on speech annotation and corpus tools*. 33(1-2): 23-60.

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber agreement for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory*, 8(2) 277-312.

Bunt, H., J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Belis-Popescu, D. Traum (2012) ISO 24617-2: A semantically-based standard for dialogue annotation. In: *Proceedings of the International Conference on Language Resources and Evaluation* (*LREC*).

Buschmeier, H. & Włodarczak, M. (2013). TextGridTools: A TextGrid processing and analysis toolkit for Python. *24. Konferenz zur elektronischen Sprachsignalverarbeitung*, pp 152–157, Bielefeld, Germany.

Gibbon, D., I. Mertins and R. Moore, eds. (2000). *Handbook of Multimodal and Spoken Language Systems: Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer Academic Publishers.

Gibbon, Dafydd (2006). Time types and time Trees: prosodic mining and alignment of temporally annotated data. In: Sudhoff, S., D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter and J. Schließer, eds. *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, 281-209.

Gibbon, D. (2011). Modelling gesture as speech: A linguistic approach. *Poznań Studies in Contemporary Linguistics* (47), 470-508.

Gibbon, D. (2013). TGA: a web tool for Time Group Analysis, in D.J. Hirst & B. Bigi (eds.) *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, 2013, 66-69.

Hirst, D. and A. di Cristo, eds., 1998. *Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Hughes, B., D. Gibbon and T. Trippel (2006). Semantic Decomposition of Character Encodings for Linguistic Knowledge Discovery. In: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, eds. (2006) *From Data and Information Analysis to Knowledge Engineering. Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Magdeburg, March 9–11, 2005*. Heidelberg: Springer, 366-373

[8] Wherever items referred to in the text are easily locatable on the internet, explicit references are generally not given.

McNeill, D. (Ed.) (2000). *Language and Gesture: Window into Thought and Action*. Cambridge: Cambridge University Press.

Mehler, A., L. Romary and D. Gibbon, eds. (2012). *Handbook of Technical Communication*. Berlin: Walter de Gruyter.

Rossini, N. (2012) *Reinterpreting Gesture as Language. Language in Action*. Amsterdam, etc.: IOS Press.

Szymański, M. and J. Bachan (2012). Interlabeller agreement on segmental and prosodic annotation of the Jurisdict Polish database *Speech and Language Technology* 14/15 (2011/2012). Poznań: Polish Phonetic Association, 105-121.

Teoh, A. and S. Chin (2009). Transcribing the Spe0ech of Children with Cochlear Implants: Clinical Application of Narrow Phonetic Transcriptions. American Journal of Speech and Language Pathology, 18(4):388-401.