LEVELS OF SEGMENTATION AND LABELLING IN THE AUSTRALIAN NATIONAL DATABASE OF SPOKEN LANGUAGE

Karen Croot, Janet Fletcher and Jonathan Harrington

Speech Hearing and Language Research Centre, Macquarie University

ABSTRACT

A database of spoken Australian English is being developed at the Speech Hearing and Language Research Centre, Macquarie University, as a preliminary to the Australian National Database of Spoken Language. The database comprises speech data which has been segmented and labelled at multiple levels, including acoustic-phonetic, broad phonetic, and intonational, with other levels to be added in the near future. This paper will describe the bases for some of these levels of labelling.

THE SHLRC-ANDOSL DATABASE

The database which has been created over the last two years at the Speech Hearing and Language Research Centre (SHLRC), Macquarie University (the SHLRC-ANDOSL database), is part of the "Australian National Database of Spoken Language" (ANDOSL), a collaborative project between four Australian laboratories (Speech Hearing and Language Research Centre, Macquarie University; National Acoustics Laboratory, Sydney; Department of Electrical Engineering, Sydney University; Department of Computer Science, Australian National University). The background to the ANDOSL initiative is described in Millar, Dermody, Harrington and Vonwiller (1990).

The SHLRC-ANDOSL database comprises roughly 4-5 hours of speech data from five adult male speakers of General Australian, each reading a total of 400 phonetically dense and phonetically balanced sentences, and a 2-minute passage, all taken from the Spoken Corpus Recordings in British English (SCRIBE) materials (Barry & Fourcin, 1992; Hieronymus *et al.*, 1990), as well as over 400 isolated words of the form /CVd/ or /CVC@d/. The data (sampled speech data, formant values, fundamental frequencies, and label files) currently occupy 2-3 gigabytes of disk space, with another gigabyte available as a "working area" for digitisation and labelling. Digitisation at 20 kHz is carried out on SUN workstations using the speech signal processing package Waves+, with Waves+ also used for hand-segmentation and labelling at the acoustic-phonetic and acoustic-intonational levels.

The speech data has been segmented and labelled by a team of transcribers, with a coordinator responsible for training the transcribers and checking the consistency of transcriptions. The levels of labelling, and the criteria for segmentation which have been used, will be described in this paper.

ISSUES IN LABELLING SPEECH DATA

Some key issues in the labelling of speech data are discussed by Barry and Fourcin (1992). Theoretically, "the "labelling" of a recorded utterance involves the temporal definition and naming of its parts with reference to the physical signal" (Barry & Fourcin, 1992, p.2). In practice, no aspect of this process is straightforward, from the decision as to what types of segment are of interest, to the question of how discrete segments can be explicitly defined within a continuous signal. As Barry and Fourcin point out, any labelling system is an abstraction from the physical signal, so the system and its conventions will be determined by the theoretical standpoint of its designers and the analytical needs of its potential users.

The first consideration is that in order to be of maximum use to all potential users, any database labelling system needs to be flexible, extendible and transparent. Flexibility allows new levels of data to be described using appropriate labelling systems, extendibility admits new phenomena within any level when these are encountered, and transparency is essential so that labelling criteria at any level can be interpreted and applied by transcribers and analysts wherever the data may be accessed (Barry & Fourcin, 1992). Among the various options for labelling possible within these broad guidelines, the SHLRC-ANDOSL database makes use of labelling levels and criteria designed very much with the

SCRIBE criteria in mind, but adapted as described below.

ACOUSTIC-PHONETIC LABELLING

In the SHLRC-ANDOSL database, the primary level of labelling applied to the speech signal is one which most closely approximates the acoustic-phonetic level described by Barry and Fourcin (1992). Our acoustic-phonetic system observes many of the principles discussed in Hieronymus *et al.* (1990) and Barry and Fourcin (1992), but it is in some ways broader and more economical, and it allows for overlapping segments.

Labelling at the acoustic phonetic level identifies events in the speech signal "in terms of established phonetic descriptors" (Barry & Fourcin, 1992, p. 3), such as stop closure, aspiration, fricative noise, nasality, glottalisation and so on. Usually one or more of these phonetic parameters provide the main cue to where to place a segment boundary. However, most segments are defined by more than one phonetic parameter, and these parameters do not always have clear cut or identical temporal boundaries. When a combination of parameters must be considered by the transcriber in establishing the boundary markers, our boundary placement follows the guidelines set down for the SCRIBE project by Hieronymus *et al.* (1990).

One goal of labelling at this level is to achieve acoustic homogeneity. This means that any two segments of the speech signal given the same label should have the same acoustic-phonetic characteristics. In practice, of course, segments given the same label may vary on any of the significant acoustic-phonetic parameters (length, periodicty, formant structure etc.), reflecting the ultimate variability, rather than predictability, of the speech signal.

However, labelling segments with reference to only the physical properties of the signal is the "physical level" of labelling described by Barry & Fourcin (1992), a level which is of limited use because it does not permit the extraction of a broad-phonetic or phonemic string from the acoustic waveform. Instead, the question of which acoustic-phonetic label to assign also depends on the phonemic identity of the segment; so that while the labels themselves do not indicate phonemic contrast, they do link the acoustic-phonetic events to the phonemic entity they represent. Table 1 shows the set of labels used at the acoustic-phonetic level in the SHLRC-ANDOSL database.

There is an obvious difference between the SCRIBE and the SHLRC-ANDOSL label sets in notation conventions used, which relates to accent differences between Australian and British English. For example, we label the vowel in "there" as [e:], because it tends to be monophthongal in Australian English, whereas in the SCRIBE acoustic-phonetic labels it is [e@], classified with the diphthongs.

A second difference is that the SHLRC-ANDOSL transcription is slightly broader and makes use of fewer symbols. The SCRIBE project has more labels to account for variation in the speech signal which is largely predictable from context. In SCRIBE, both the closure and release phase of oral stops are differentiated according to place of articulation (thus [bo bb] and [to tb] for voiced bilabial and voiceless alveolar stops (Barry & Fourcin, 1992)), whereas at SHLRC we use the same symbol, [H], for the release phase of all stops (hence [b H], [t H]). We also do not separately label retroflex versus non-retroflex schwa, the labialised portion of fricatives, nor nasalised or fricated sections of vowels. This sort of variation is time-consuming to transcribe and difficult to transcribe consistently. Where necessary, such "predictable" variation may be studied empirically at a later time by extracting, according to context, and acoustically analysing, all segments potentially containing the type of variation in question. For similar reasons, we do not currently differentiate between the voiced and unvoiced parts of stop closures, fricatives, nasals, liquids, glides and vowels.

The SCRIBE system also has a large range of symbols for non-speech fillers such as breath noise, tongue click and lip smack, whereas we include all pauses and pause-associated sounds in the one label [#]. Finally, and also in order to minimise the complexity of the label string, creaky and breathy voice are not marked at the acoustic-phonetic level but will be included in a separate "laryngeal level" to be added to the database soon. Creaky voice which is associated with glottalisation between, before, or during vowels is labelled at the acoustic-phonetic level.

Although we have tried to minimise the labelling of "predictable" variation, there is always the question of what variation is predictable from context and what is not. For example, most retroflexion versus fronting of schwa is due to context. However, it is not always clear whether the variation in some unstressed vowels represents an allophone of schwa or an allophone of a full vowel. Thus, in Australian English, the words "explain", "entrance" and "again" may be perceived as beginning with [I], [E] or [V], or with schwa. In the final instance, the choice of label assigned to a particular segment depends on the transcriber's perception of that segment in context.

Apart from selecting an appropriate label for a segment, the other main responsibility of the transcriber is to decide whether a segment which might be expected to occur has actually been produced in the spoken data. Elision and assimilation in continuous speech all too commonly produce a spoken form which bears limited resemblance to the abstract phonemic form which might be predicted from the orthographic form of the utterance. Transcribers decide whether a segment is present or not, or assimilated to another segment or not, by listening to the word or phrase in which the segment occurs, as always relying on their percept of the segment in context and referring to the acoustic-phonetic cues on the signal displays in Waves+ (spectrogram, time-pressure waveform and f0 display).

In transcribing continuous speech, any of the following may occur:
i)the segment is not perceptible at all (eg. [o:weiz]),
ii)the segment is perceptible but there are no cues on the speech waveform or spectrogram to indicate where the boundaries should be placed (geminates, adjacent fricatives etc.), or
iii)the transcriber is unsure as to whether the segment is perceptible in its own right or is assimilated to an adjacent segment.
In the case of (i), the segment is assumed to be elided and therefore is not labelled as present in the signal. Where there is uncertainty about assimilation, due to a lack of acoustic-phonetic cues or perceptual cues (ii) and (iii), a double label is used, in which both appropriate acoustic-phonetic labels are assigned to the one segment of the speech signal. In subsequent analyses, the broad-phonetic segments corresponding to the two acoustic-phonetic segments in the double label are defined as "temporally overlapping" and given the same boundary times (McVeigh & Harrington, this volume). This broad-phonetic level of labelling will now be considered in more detail.

BROAD-PHONETIC LABELLING

We have already referred to the need to be able to derive a phonemic-based string from the acoustic-phonetic labels. Such a string provides a bridge between the acoustic signal and the lexicon, necessary for any subsequent analysis of the data. Like Barry and Fourcin (1992), we have rejected a purely phonemic string as impractical because phonemic labelling does not take into account the characteristics of continuous speech, and because many utterances have more than one possible citation-phonemic form. Also like Barry and Fourcin, we have therefore opted for a broad-phonetic string as the required "bridge".

In a broad-phonetic level of labelling, "speech sound symbols that have phonemic status" are used "to indicate non-phonemic i.e. continuous speech phenomena" (B&F 1992. p.10). The broad-phonetic string has the advantage of retaining the continuous speech phenomena of assimilation and elision, while avoiding the theoretical issues as to what is an "appropriate" phonemic representation. In the SHLRC-ANDOSL database, the broad-phonetic label string is derived automatically using a recursive parsing strategy based on the orthographic form of the utterance (McVeigh & Harrington, this volume).

Barry and Fourcin (1992) point out that at this level there is greater phonetic and acoustic variation across segments given the same label, rather than homogeneity of segments. For example, at our broad-phonetic level, events including closure, release phase and/or glottal stop are all collected under the broad-phonetic label for any plosive. However, this is an unavoidable and necessary artefact of mapping a limited set of linguistically meaningful symbols onto an infinitely variable physical signal, and codifying this variability is an essential task of the speech scientist.

ACOUSTIC-INTONATIONAL LABELLING

At this level, pitch accents and intonational boundary tones are marked as events on the changing

fundamental frequency contour. Our system for acoustic-intonational transcription follows that of Pierrehumbert (1980), Pierrehumbert and Beckman (1988), and more specifically that of the TOBI (tones and break indices) system which is currently used on a national project in the U.S. for prosodically annotating speech data.

There are three distinct hierarchical levels at which intonation can be marked. The first of these is the "pitch accent" level, in which every accented syllable is assigned a pitch accent tone. The basic accents are the peaked accent H*, corresponding to a tone target in the speaker's upper and mid pitch range, and low accent L*, corresponding to a tone target in the lowest portion of the speaker's pitch range. These tones can be modified with diacritics to indicate a rising peak accent L+H*, scooped accent L*H, and downstep (indicated by ! diacritic).

The second level at which tones can occur is the "intermediate phrase". An intermediate phrase is a grouping of words which includes at least one pitch accented syllable: tones at this level are either low, L, or high, H, and are known as "phrase accents". The pitch range is reset at intermediate phrase boundaries, therefore downstep on pitch accented tones only ever occurs within an intermediate phrase (Beckman & Pierrehumbert, 1986). Phrase accents may themselves be marked for downstep in certain circumstances.

The third tonal level is the "intonation phrase", which is a grouping of intermediate phrases. Tones at this level are known as "boundary tones", and can be either low, L%, or high, H%. Phrase accents are combined with boundary tones to produce four major boundary pitch contours: falling declarative, L-L%; continuation rise, L-H%; mid-level tone, H-L%; and high rise H-H%.

CONCLUDING REMARKS

The database is currently being extended to include articulatory data (movement parameters from the lip, jaw, tongue and facial muscles, airflow, air-pressure and laryngeal waveform data) from the five male speakers, as well as hand-segmented data from five female speakers. Data which has already been labelled is being checked for consistency, using the mu+ system developed at SHLRC (McVeigh & Harrington, this volume) for matching multiple labels to signal files in a single display. Although we have sought from the beginning to economise by not labelling some aspects of the physical signal as discussed above, our present set of labels has been increased from an initial minimal set in order to cover the phonetic features that we have observed. Thus, we have had to add labels for fricated stop closures, silences between fricatives and nasals, bursts on liquids and nasals, trilled r etc., whereas the maximal label set used for the SCRIBE project would have ensured these phenomena labelled from the start. However, the labelling system is extendible in view of just such a need, and the advantage of using the more economical set of labels at SHLRC has been quicker labelling, easier and quicker training of transcribers, and greater consistency and reliability of transcription.

REFERENCES

Barry, W. J. & Fourcin, A. J. (1992) Levels of labelling. *Computer Speech and Language, 6,* 1-14.

Beckman, M.E. & Pierrehumbert, J.B. (1986) Intonational structure in Japanese and English. *Phonology Yearbook, 3,* 255-310.

Hieronymus, J., Alexander, M., Bennett, C., Cohen, I., Davies, D., Dalby, J., Laver, J., Barry, W., Fourcin, A., & Wells, J. (1990) Proposed speech segmentation criteria for the SCRIBE project. SCRIBE Project Report.

Millar, J., Dermody, P., Harrington, J. & Vonwiller, J. (1990) A national cluster of spoken language databases for Australia. *Proc. of the Third Australian Int. Conference on Speech Science and Technology.* Melbourne, Australia.

Pierrehumbert, J.B. (1980) The Phonology and Phonetics of English Intonation. Ph. D. dissertation, M.I.T., Cambridge, Ma. (Distributed by the Indiana University Linguistics Club, Bloomington).

Pierrehumbert, J.B. & Beckman, M.E. (1988) *Japanese Tone Structure.* M.I.T. Press: Cambridge, Ma.

Oral Stops

Closures:	[p], [t], [k], [b], [d], [g]
Incomplete closures:	[pH], [tH], [kH], [bH], [dH], [gH]
Glottal stops (substituting for stops):	[p^], [t^], [k^], [d^]
Release (burst and/or aspiration):	[H]

Affricates

Closures:	[t], [d]
Frication:	[S], [Z]

Fricatives [f], [v], [T], [D], [s], [z], [S], [Z], [h]

Sonorants

Nasals:	[m], [n], [N]
Syllabic nasals:	[=m], [=n]
Voiceless nasals:	[Om], [On], [O=m], [O=n]
Release (burst and/or aspiration):	[mH], [nH], [NH]
Approximants:	[w], [j], [l], [r]
Syllabic approximants:	[=l] (deletion site before [l]) [l=] (deletion site after [l])
Voiceless Approximants:	[Ow], [Oj], [Ol], [Or]
Release (burst and/or aspiration):	[lH], [rH]
Trilled /r/:	[r"]

Long vowels		Short vowels	
[i:]	heed	[I]	hid
[u:]	who'd	[U]	hood
[e:]	hair	[E]	head
[@:]	heard	[@]	the
[o:]	hoard	[O]	pod
[a:]	hard	[V]	hut
		[A]	had

Diphthongs

[ei]	say
[@u]	go
[oi]	toy
[ai]	high
[au]	how
[i@]	here
[u@]	cure (occasionally)

Glottal stops and glottalisation between vowels
indicated by symbol for adjacent vowel followed by C
eg. [g l u: u:C u: z d H] glue oozed

Miscellaneous

Start of utterance:	H#
Pause and/or speech fillers:	#

Table 1. Labels used at the acoustic-phonetic level in the SHLRC-ANDOSL database.