Proceedings of the Institute of Acoustics

STANDARDS FOR THE PHONETIC LABELLING AND SEGMENTATION OF TELEPHONE QUALITY SPEECH

C M Scott(1), R Lickley(1), K Edwards(1) & A Simons(2)

(1) Centre for Speech Technology Research, University of Edinburgh
(2) BT Laboratories, Martlesham Heath

1. INTRODUCTION

This paper presents the labelling criteria employed for the phonetic segmentation and labelling of a telephone quality speech database (Subscriber). The data was collected by BT Laboratories and annotated at the Centre for Speech Technology Research using the Centre's Speech Labelling Workstation, or SEGSTAT. The workstation was designed for earlier speech labelling work and was customised for use in the Subscriber project. Labellers work on time-aligned speech waveforms and spectrograms using a mouse to playback, define segment boundaries and assign symbols and diacritics from 'pop-up' tables. Subsequent reallocation of boundaries, relabelling and division or merging of segments is also possible. The IPA based symbols and diacritics which appear on SEGSTAT and in this paper have corresponding ASCII representations which appear in the finished transcription files alongside a start and stop time in samples for each segment. These corresponding machine-readable symbols were developed at CSTR.

The purpose of the Subscriber [1] database was to create a detailed archival reference of the telephone speech of some 1000 speakers from around the United Kingdom. The segmentation was carried out at a detailed acoustic-phonetic level, where labels incorporated not only the actual speech but also any incidence of telephone line noise, background noise from the caller site or extra speaker noise. The range of possible speaker accents and the fine prescription of the acoustic-phonetic labelling led to the adoption of an 83 member symbol set plus 17 diacritics to cover 9 British accent groups. Annotations and quality control standards were defined during the Bader [2] database pilot study.

In addition to the annotation, each speech file was assigned a 'header' containing further relevant information such as an accent decision made by the labeller. The accent decision for each speaker was made on the basis of two of the 7 sentences which each speaker uttered. These were specially designed 'Shibboleth' (accent-determining) sentences [3]. There was a total of 9 accent groups to choose from, allowing the labeller to describe any accent in the United Kingdom using any combination of the accent groups, for example, Southern British Standard (RP) with Scottish influence.

2. STANDARDS FOR SEGMENTATION AND LABELLING

2.1 Introduction

Previous speech labelling work carried out at CSTR (SCRIBE [4], ATR [5]) has provided the Centre with a wealth of experience in the annotation of wide-band speech. This resource was the starting-point for the creation of the segmentation and labelling standards for Subscriber.

Since the Subscriber database was collected from various sites in the UK, a range of regional accents are represented and a comparative phonetic transcription was required to separate different pronunciations of systemically similar vowels and consonants. The transcription also includes certain acoustic events, allowing sub-phonemic analysis of the data.

The symbol and diacritic set employed in the labelling of Subscriber has its foundation in IPA. It is augmented with other special characters adopted as symbols and diacritics to label noise and other phenomena. These are detailed below. The vowel and consonant symbols in the set were chosen with reference to Wells [6] account of the accents of English and other works. Only those relevant to the accent areas represented in Subscriber are included.

2.2 Difficulties in Labelling Telephone Quality Speech

There are some common characteristics of telephone quality speech that make it more difficult to label than wide-band speech.

Firstly, there is a lack of voicing information, making it almost impossible for the labeller to tell whether fricative, closure or burst segments are voiced or voiceless. In these cases, contextual information is used in order to identify the segment, although enough information is usually present to allow an accurate segmentation to be made. (The two diacritics v and o, representing voicing of a voiceless sound and devoicing of a voiced sound, respectively, were included in the diacritic set, for use with all appropriate symbols, although in practice they were rarely employed).

The narrow bandwidth of telephone speech limits the amount of high frequency information visible in the spectrogram, often resulting in fricatives appearing as a blank section, resembling a stop closure phase. This is particularly true of voiced fricatives where the frication is inherently not so strong. Accurate boundaries can still be drawn using contextual information, although the end boundaries of utterance-final fricatives must often be placed using auditory information.

Weak or poor signals are a common source of difficulty in labelling telephone quality speech. There may be very little information, either in the waveform or the spectrogram to help distinguish the segments, and auditory information has to be relied upon. Weak signals may be caused by the caller talking too quietly or too far from the handset or through some aspect of the network. The problem of weak signals appears to be more common with female speakers, not just because they are more likely to have quieter voices but also because the formant structure of female speech is not as strong and clear as that of male speech in the best of conditions.

Finally, the noise that is often present in telephone speech can cause problems for the labeller. Noise may be due to the line or the speaker or it may be backround noise from the caller site. The noise categories are discussed in more detail in section 2.9. Although many noises are very short term events and do not significantly distort the underlying speech, others can mask it completely for large sections of the speech. Such events affect the quality decision made on the file, see section 3.

2.3 Stops

[p b t d k g ʔ]

Where possible, the different phases of stop consonants (closure plus burst in voiced stops, and closure followed by burst and aspiration in voiceless stops), are labelled separately in the database. If the signal is very weak or noisy or there is very heavy frication, making the stop closure indistinguishable from the following release, the phases are collapsed and the stop is labelled as one event, using the relevant stop symbol.

Where a closure phase is clearly distinguishable from the rest of a stop, usually appearing as a blank portion on a spectrogram, it is given a label consisting of the stop symbol plus a □ diacritic, representing closure. The diacritic, ᶠ, indicating frication, can be added to a stop closure label where relevant. Stop closures occuring in utterance-initial or final position, which offer no information on the spectrogram or waveform which could lead to a positive labelling decision, are marked with an arbitary closure of 50 msecs. Where a sequence of two stops occurs (e.g. in 'directors') and the first stop remains unreleased, the long closure phase is split equally between the first and second stops. The same applies to geminate stops (e.g. in 'The bulb blew...') which are split into two equally sized tokens of the same stop.

Where it is possible to identify a clearly defined burst release phase of a stop, appearing as a relatively dark portion on the spectrogram, it is labelled with the appropriate stop symbol plus a ! diacritic, indicating a burst release. In stop plus fricative clusters, where the release phase of the stop cannot be distinguished from the following fricative, no burst is marked and the fricative immediately follows the stop closure. In stop plus nasal clusters where there is no burst release and the stop closure is released nasally by the lowering of the velum, the relevant nasal symbol immediately follows the stop closure.

Where a fricative-like aspiration phase follows the release of a voiceless stop, it is labelled with the stop symbol plus the diacritic ʰ, representing aspiration. Sometimes a release burst may be weaker than normal and indistinguishable from the following aspiration. In such cases the two phases are collapsed into one category and given the stop symbol plus ʰ diacritic as a label. In words like 'cucumber', where the aspiration phase of the voiceless stop (here the initial /k/) contains all the quality information of the approximant (/j/ in this example) and the following voiced section contains only the vowel, it is desirable to indicate the presence of the approximant. Thus, the section of aspiration that contains its particular quality is given the approximant symbol plus the ᶠ (fricated) diacritic.

In some accents, particularly Southern British Standard, a glottal stop commonly occurs after a vowel or approximant to reinforce [7] a following voiceless stop. This glottal stop is labelled as such, using the ʔ symbol, and is not treated simply as part of the stop closure phase. The ʔ symbol is used wherever glottal stops occur.

2.4 Fricatives

[ʍ f v θ ð s z ʃ ʒ x ɣ h]

As mentioned in 2.2, the characteristic 'snowflake' appearance of fricatives on a spectrogram is largely lost and many, especially those which are voiced, appear as blank portions. Accurate boundaries can be drawn and labels assigned using contextual information. Epenthetic silence, a common feature of fricatives in conjunction with vowels, nasals and approximants, is not given a separate label in the Subscriber database but is labelled as part of the fricative. The onset and offset of the fricative is taken as the point where visible energy in the formants of the surrounding vowels, nasals or approximants ceases or begins. Auditory information must often be used to place the final boundary of utterance-final fricatives.

2.5 Affricates

[ʧ ʤ]
In common with the labelling policy adopted for stops, the different phases of affricates are labelled separately where possible. The closure phase is marked with the appropriate affricate symbol and the closure diacritic. The release of affricates involves an initial release burst followed by strong frication. Often the burst is indistinguishable from the frication and the whole portion is labelled with the affricate label alone. Where a separate burst is obvious it is labelled with the affricate symbol plus the burst diacritic and the following frication is given the affricate symbol alone. Where an affricate's closure phase is fricated but still distinguishable from the following frication it is marked as a closure plus the diacritic f. However, if no separate closure phase is discernable in the frication then the affricate symbol alone is used for the entire segment.

2.6 Approximants

[ɫ l ƚ w j ɹ]
The two allophones of /l/, 'clear' or palatised /l/ - [l] and 'dark' or velarised /l/ - [ɫ], are labelled separately in Subscriber. Intervocalically, onsets and offsets marked by rapid formant transitions make segmentation reasonably easy and clear /l/ is characterised by a higher F2 than dark /l/. In other environments, however, transitions are not so obvious and onset and offset is estimated using auditory clues. The 'vocalised' dark /l/ that exists in some accents, for example that of some speakers from the London area, is also labelled separately in Subscriber using the ƚ symbol. Syllabic /l/ is marked with the syllabic diacritic, =.

The onset/offset of the approximants [j] and [w] is again estimated. This is particularly difficult intervocalically but is identified as the midway point between a strong percept of the vowel context and a percept of the approximant.

Two allophones of /r/ are labelled separately in the Subscriber database. These are the approximant, [ɹ] and the tap, [ɾ] (see 2.6.1 below). The approximant /r/ is isolated from its context using the same approach mentioned above for other approximants.

2.6.1 Tapped /r/

[ɾ]
By its nature the tap is easily isolated from its context. It appears like a voiced stop of very short duration, making boundary placement straightforward.

2.7 Nasals

[m n ŋ]
Onsets and offsets of nasals are identified by the characteristic discontinuity in formant frequencies and the reduction in formant amplitudes. Nasal sequences are relatively easily divided due to the different formant frequencies of each nasal. However, in telephone quality speech this information is often lost or severely reduced in the case of nasals, due to a weak or distorted signal. Where

division is impossible, the same approach that applied to sequences of stops is adopted and the nasal portion is split into two equal parts which are given the appropriate labels. Where an utterance-final nasal exhibits a burst release this is marked separately from the rest of the nasal using the appropriate nasal symbol plus a burst diacritic. Syllabic nasals are marked with the syllabic diacritic.

2.8 Vowels

i ɪ e ɛ æ a u ʊ o ɔ ʌ ɑ ɒ ɜ ə ɝ ɚ]

[iə ɪə ɪi ɪu ei eə eʊ ɛə ɛɪ æʊ aʊ ai aɪ ae ɑɪ ʊə ou ɔə ɔe ɔɪ ɔʊ ʌʊ ʌi ɒʊ ɒɪ əi əɪ əʊ]

Segmentation of vowels in an environment of stops and fricatives presents the labeller with few problems. In other environments, the labeller must rely heavily on auditory clues. In the Subscriber database, diphthongs are not split into their component vowel qualities as they were in the SCRIBE database [4]. Instead, each diphthong has its own symbol. The diacritics for nasalisation (~), lengthening (ː), frication (ſ), and glottalisation (ʔ) qualify the vowel symbols where necessary. The vowel set created for Subscriber proved adequate for the purposes of labelling the database. Some vowel symbols were widely used, others rarely. However, no obvious deficiency was discovered in the course of labelling the database.

Stress was assigned at a lexical level in Subscriber, using the diacritics ' and " to represent primary and secondary stress respectively.

The diacritic level of detailed information on stress, glottalisation, frication etc, allows the level of phonetic detail included in Hidden Markov Models to be varied.

2.9 Noise and other Special Characters

[# % < ₩ ¦ --- ?-]

These special characters were developed for the Subscriber database to enable labellers to categorise the different types of noise encountered in the speech files.

The # label indicates a 'silent pause' and is used to mark any section of a speech file that does not contain any speech or noise (epenthetic silences and stop closures excepted).

The 'breath-filled pause' label, %, is used where noise of the speaker breathing can be heard or seen in the signal. This includes pre-utterance inhalation and post-utterance exhalation.

The < label is used for any other speaker noise outwith the actual speech. This includes things like coughs, lip-smacks and tongue clicks. However, it does not include any extra speech on the part of the speaker, which is marked using the ₩ symbol. This symbol is also used when extra speech produced by a person other than the speaker can be heard. Where extra speech produced by the speaker can be labelled, it receives the appropriate labels plus the extra speech diacritic, ⁄, throughout.

Where a single impulse noise occurs during a period of silence or pause, whether it be background noise from the caller site or noise resulting from the telephone line, it is marked with the label ! . If such a noise occurs during the actual speech, an impulse noise diacritic, ¦ , is used. Where there is a series of impulse noises, the label --- is used to indicate periodic impulse noise or the corresponding --- diacritic is used if appropriate.

A final noise label, -?- , is used to label any 'unidentifiable' noise outwith the speech. This includes line noise which cannot be labelled with the impulse or periodic impulse symbols and any speaker noise or background caller site noise which cannot be labelled using any of the identifiable noise labels. A diacritic -?- is also used where appropriate.

A final diacritic to be mentioned is /. It is used to highlight disfluent sections of speech (including 'um' and 'er' type hesitations), which are fully labelled according to their phonetic content. This diacritic is also added to any post-disfluency pause. The occurrence of disfluencies is noted in the header of the file in question and will of course have an effect on the quality decision (see below).

3. HEADER FILES

Several important pieces of information supplement the label data given for each speech file and are given in a header file. The name of the corresponding speech file, the sampling frequency, the length of the file in seconds, the text of the utterance, an accent decision, a quality decision, the name of the labeller, the date and any other relevant comments are all included.

Perhaps the most important entries in the header are those of 'text', 'accent', 'quality' and 'comment'. The actual contents of the file are entered in the 'text' field, carefully noting the exact utterance as well as any deviations from the prompt and any disfluencies or cut-offs. The accent decision for the speaker, made by the labeller based on the accent-determining clues given in the Shibboleth sentences, is inserted under 'accent'. This decision can be further qualified in the 'comment' field, for example noting non-rhoticity with a rhotic accent decision or influence from one or more other accents. Other relevant information such as indication of disfluency and mention of noise or weak signal can be entered under 'comments'.

Finally, a quality decision made by the labeller about the utterance is included. A 'good' utterance' is one which follows the prompt exactly and is not marked by any significant line noise, background noise or distortion. A 'usable' utterance may differ from the prompt but is otherwise good (including those with disfluent sections or which are cut off), or it may be marked by line or background noise but is still labellable. A 'bad' utterance is one that is so badly articulated, or produces such a weak signal, that it is very difficult to label with any degree of accuracy.

4. THE ACCENTS

The Subscriber database contains speech collected from around the British Isles, encompassing several accents of English. They fall into 9 categories which are abbreviated in the table below.

SBS	Southern British Standard
LON	London Area
R-WEST	West of England (Rhotic)
WAL	Wales
NB-LIV	Liverpool Area
NB	North of England
R-LANCS	Lancashire (Rhotic)

R-IRISH	Ulster (Rhotic)
R-SCOTS	Scotland (Rhotic)

The labeller makes an accent decision for each speaker based on the Shibboleth sentences. The sentences were designed to contain at least two clues for each accent. A full account of the design of the Shibboleths and the basis for accent decisions is presented in [8]. The two sentences are given below.

1. I hear that some young bears were caught off guard by the rising tide.
2. One of the cubs soaked its brown pelt in a bath of muddy grey sea-water.

The labeller's accent decision notes which of the 9 accent group specifications best describes the speech of each individual speaker, based on observable phonetic qualities of vowels and consonants. Where a speaker's accent influences are mixed, a firm decision is still made but it is qualified with a comment in the header, as described above.

5. CONCLUSIONS

The phonetic labelling standards adopted for Subscriber led to the data being annotated at a detailed, acoustic-phonetic level. This included not only actual speech events but also phenomena peculiar to telephone speech, in particular noise. The symbols and diacritics adopted for the labelling task proved adequate to cover the range of phonetic qualities and types of noise encountered. Some symbols were more utilised than others, but there were no obvious gaps in the range of descriptive labels.

The level of detail of the finished database and the nature of the phonetic description provide BT with a unique and highly useful resource for use in their speech research.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge the support given by BT Laboratories in funding this research.

7. REFERENCES

[1] A Simons & K Edwards, 'Subscriber - a Phonetically Annotated Telephony Database', *these proceedings* (1992)
[2] K Edwards, C M Barr & R Lickley, 'Proposed Speech Segmentation Criteria for the Bader Database', *CSTR internal document* (1991)
[3] J Laver, 'Accent-Diagnostic Indicators in British English: Implications for Automatic Speaker-Independent Speech Recognition', *Subscriber Project Report* (1990)
[4] J Hieronymus, M Alexander, C Bennett, I Cohen, D Davies, J Dalby & J Laver, 'Proposed Speech Segmentation Criteria for the SCRIBE Project, *CSTR internal document* (1990)
[5] J Laver, M Alexander, C Bennett, I Cohen & D Davies, 'Speech Segmentation Criteria for ATR/CSTR', *CSTR internal document* (1989)
[6] J C Wells, *Accents of English* (Volumes 1 and 2), Cambridge University Press, Cambridge (1982)
[7] J C Wells, 'A Phonetic Update on RP', *Speech, Hearing and Language, Work in Progress*, Volume 5 (1991)
[8] K Edwards, J Laver, M Jack & A Simons, 'The Design and Performance of Two Accent Diagnostic 'Shibboleth' Sentences', *these proceedings* (1992)