# PROBLEMS OF TRANSCRIPTION AND LABELLING IN THE SPECIFICATION OF SEGMENTAL AND PROSODIC STRUCTURE

Martine Grice and William Barry

Department of Phonetics and Linguistics, University College London

## ABSTRACT

A consensus transcription by two independent phoneticians of four speakers' readings of a two-minute passage was compared with the hand-labelling of the same recordings. A broad phonetic level of transcription was employed as this level of representation is usual in speech technology applications; the same symbol inventory was used for labelling. A number of differences between transcription and labelling are discussed with reference to the theoretical problems of relating auditory symbolic representation to the acoustic signal; the mapping of transcribed elements onto temporally defined acoustic segments is less than straightforward.

## 1. INTRODUCTION

There is a long tradition behind a number of internationally established conventions for the auditory specification of the segmental structure of utterances. In the case of prosodic structure, the conventions are of a more language-specific and theory-bound nature. With the growing availability of digitised speech recordings, facilities for representing the speech signal graphically, and the urgent need for large annotated speech databases, these conventions are being challenged. Labellers are faced with the task of relating two phenomenologically different manifestations of speech: a symbolic representation using transcription symbols and a two dimensional transformation of the physical speech signal on the screen.

This paper addresses this theoretical dilemma. The cases of segmental structure examined include a)voiceless schwa b) the occurrence of glottal stop or constriction and c) linking and syllabic /r/. Though these are only a small number compared to the speech sounds transcribed and consistently segmented and labelled in these recordings, they are the product of normal articulatory processes, not freak events which could be safely ignored in the description of normal continuous speech. While they disappear as "noise" in the training distributions of stochastically based speech-recognition systems, they are in fact signal properties which, consistently labelled, could provide additional structural information for the generation of phonological rules. The prosodic investigation focusses on tone-group demarcation in the transcription: the consistency with which transcribers place tone-group boundaries and the relationship of these boundaries to labelled pauses.

## 2. SPEECH MATERIAL AND ITS REPRESENTATION

For the segmental analysis, four recordings of approximately two minutes each of the "Numbers Passage" from EUROM.0, the first CD-ROM database of the Esprit Project 2589 (SAM) were transcribed by two trained and experienced phoneticians. They were also hand-labelled by a number of SAM research assistants and cross-checked by the authors.

The level of detail given in the auditory transcription can be described as "broad phonetic". Only symbols available from the phonemic inventory of Southern Standard English were used, but changes to the phonemic structure of words resulting from continuous speech processes were captured, e.g. [k@m bi] for "can be" or [b{g k@Uld] for "bad cold". On the other hand, similar assimilatory processes such as dentalisation of /n/ in sequences such as "on the" or "in that" are not distinguished from regular alveolar realisations. Syllabic sonorants were represented (e.g. [=m, =n, =l]).

The same inventory was employed for labelling which was based on visual scrutiny of the speech pressure waveform and simultaneous auditory examination of sections no shorter than one syllable in length.

## 3. TRANSCRIPTION VS LABELLING

### 3.1. Segmental issues

The discussion concentrates on those aspects of symbolic description which are sensitive to a) the difference in size of processing frames generally available in auditory transcription compared to those used at a labelling workstation, and b) the discrepancy between the transcription goal of merely perceiving a string of appropriate sounds and the need in labelling to annotate every part of the signal and to associate each element in the symbol string with a discrete signal segment.

*a) Voiceless schwa?*

The possible elision of schwa in extremely reduced syllables is well accepted for English (e.g."proprietary" ending in [t@ri] or [tri] [2]), but it is not usually given as a possible process in the weak form of "to". However, in the many occurrences of the phrase "to the", there were several cases of the /t/ release merging into the following interdental fricative /D/. Perceptually, the impression was of two, albeit extremely short, syllables, so the Broad Phonetic transcription was, logically, [t@D@]. The labelling decision was just as clear: in the case of stop-vowel sequences, the release burst and associated following frication are normally counted as part of the stop, periodicity in the signal is necessary for the labelling of a vocalic segment. Consequently, the absence of a part of the signal on which to attach a vowel label led to the labelled sequence [tD@].

Although a narrower phonetic transcription might have shown the schwa to be voiceless, it is doubtful whether, given the disyllabic percept, and the occurrence of a true vowel very soon after, that the voicelessness of the first syllable can be perceived in the normal flow of the phrase. Only by isolating the two syllables with a speech editor does this become apparent. The artificiality of the segmentation criteria is readily apparent; post-release aspiration, however short, has the function of identifying both a vowel and a consonant. In contrast, devoicing of liquids in stop-liquid sequences is a generally accepted phenomenon (e.g. in "place", "try"). Here transcribers and labellers can be in complete agreement: periodicity is not required, so the frication following the stop burst may be labelled as part of the liquid.

*b) Glottal stop or constriction?*

The glottal stop is not part of the phonemic inventory used for transcription and labelling. However, glottal stops are clearly audible in the recordings. It is well known that the glottal stop has a number of functions in English. It can reinforce voiceless plosives, replace /t/, and mark the vocalic onset of a stressed syllable. These three different functions correspond, at least in theory, to different manifestations in the speech signal: Reinforcing /p, t, k/, it occurs just before the period of silence resulting from the stop closure [2]. In both of the other two cases it can occur between vowels; glottal onset occurs at the beginning of *stressed* syllables and glottal /t/ replacement alone only occurs before *unstressed* vowels. The combination of glottal /t/ replacement + glottal vowel onset is differentiated from glottal vowel onset alone by the checked nature of the preceding vowel.

Thus, perceptually there are no problems in identifying them, and in a narrow transcription task their occurrence could be recorded. In a labelling task, however, they present manifold problems, since they cannot be treated consistently. A perceived glottal stop may consist of a period of total glottal closure similar to a stop closure, but more frequently it is characterised in the speech signal by either of the following:

i) a rapid increase in the duration of the laryngeal period, and, after a number of very long periods, a return to normal duration;

ii) a reduction in initial peak amplitude over one or two cycles, often accompanied by irregular period duration.

We use the cover term "glottal constriction" for the above two cases when there is no "glottal stop" as such.

Preceding a period of silence, lengthened or irregular glottal periods can arguably be considered part of the preceding vowel or sonorant (since the spectral properties are unchanged). They may, however, signal the preparation for a glottal onset of the following vowel (and are therefore a cue to the ensuing glottal closure), the reinforcement of a stop consonant (the closure phase of which constitutes the silence) or the replacement of a /t/ (where the silence is due to a glottal closure). All three can be auditorily distinguished and do not pose problems for the labeller at the broad phonetic level since the only symbol type permissible is a plosive which can be conveniently attached to the silence; the glottal reinforcement of vowels does not have to be marked.

In the case of /t/ replacement, there are problems with glottal constriction; the amplitude and period irregularity in the signal may constitute the only part of the signal which can be associated with the /t/. Furthermore, it does not need to be very accurately placed to result in an acceptable percept. It may be regarded as an "overlay" on a slowly changing vocalic or sonorant sequence, and it does not always occur in the transitional phase. In one extreme case, "point zero" was produced with the glottal /t/ replacement occurring at the *beginning* of the [n] segment (see figure 1). The most accurate labelling of the event was [poIt=n], although this aberrant sequence was only perceivable after isolation with the speech editor.

*c) Linking and syllabic /r/*

/@r/ and /r@/ sequences resulted in a number of transcription-labelling divergencies. They were consistently transcribed as diphonic, while in seven cases they were labelled as [=r]: in "numerals", "score and" and "number eight". The case of "numerals" is similar to that of words such as "preference" and "different", given in the literature [3] where compression can also occur. Two labelled versions were found: [njum=r@lz] and [njumr@lz], depending on whether two or three syllables were perceived. Syllabic /r/ is not generally said to occur as a manifestation of linking /r/. In "score and" and "number eight" it was found for both /r@/ and /@r/: [skO:=rn] and [nVmb=r eIt]. The nature of schwa and /n/ make their segmentation in the speech signal arbittrary, even when a schwa is clearly present. In the seven cases recorded as [=r], there was no schwa portion distinguishable from the /r/, and the /r/ was confirmed as syllabic by listening to it with only the preceding context (using the speech editor). Listening to the syllable sequence in context (transcription mode) reduced the clarity of the [=r] percept.

### 3.2. Tone-group demarcation

Common to most systems of intonational representation is the provision for intonational boundaries of some kind. Within the British approach, such units consist of the tone-group, containing an obligatory nucleus. Since the only prosodic labelling of Eurom-0 presently available is the specification of pauses, the consistency across transcribers in placing these boundaries and the relationship of these boundaries to labelled pauses were examined. To this end, eight experienced British phoneticians were asked to transcribe tone-group boundaries in the first two paragraphs of one speaker's rendition of the numbers passage.

Of the 43 boundaries which appeared in the transcripts, 32 were unanimously transcribed. 26 of these coincided with

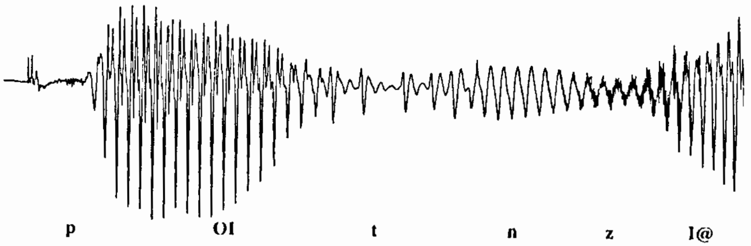

Figure 1 Glottal replacement of /t/ in the phrase "point zero"

pauses (>50ms) in the signal and six had complete agreement without a signal pause. Discrepancies in the transcriptions were found at 11 locations (none with signal pauses).

Four cases of disagreement occurred in the sequence "two, three, four, five, six, seven, eight, nine". Its tonal structure can be informally described as an alternating sequence of relatively high and relatively low shallow rises of the type:

tw$^{o}$ thr$^{ee}$ fou$^{r}$ fiv$^{e}$ si$^{x}$ sev$^{en}$ etc.

Two of the transcribers grouped together sequences of "high" rises followed by "low" rises and six marked a tone-group boundary between each item in the sequence.

The tonal and rhythmic patterns in the sequence suggest that there is a need for capturing a hierarchical structure beyond that of tone groups and intonational paragraphs. The list of digits might most satisfactorily be seen as a pattern of minimal, single-word units, co-ordinated, in turn, within larger ones. Whether the single-word units are below that of the tone group (cf Beckman and Pierrehumbert's intermediate phrase[1]) or above, is an issue for further theoretical discussion. What matters is that, for the transcription of continuous speech, a variety of prosodic boundary types might engender greater consistency amongst transcribers, a necessary prerequisite for the investigation of correspondences between transcribed boundaries and properties of the signal.

It is clear that the elements in the sequence are syntactically linear since they are components of a simple list. It would be interesting to ascertain whether transcribers would be more consistent in placing boundaries when a syntactic hierarchy is apparent such as in the phrase "red book, green pen, brown desk" where a similar tonal pattern might be produced:

re$^{d}$ boo$^{k}$ gre$^{en}$ pe$^{n}$ brow$^{n}$ des$^{k}$

A second area of inconsistency was in the case of a) "be composed", b) "no system" and c) "keep pace" A boundary was marked by five transcribers in a) and b) and by four in c). In a) and b), there were also discrepancies between the transcriptions of accented syllables. In c), all were transcribed keep pace. In all three cases, there was a high pitch on the first word and a low level pitch on the second as follows:

a)$^{be}$ composed b)$^{no}$ system c)$^{keep}$ pace

Transcribers who recorded a boundary after the second word also marked it as accented. Some transcribers may have been reluctant to place a boundary in these positions because the pitch on the second word was level, this being the most problematic of the tones in a theory where dynamic pitch is generally seen as an indicator of nuclearity.

A hierarchical intonational structure might facilitate the task of the transcriber here too. The intonation of the phrase in which b) and c) appear: "no system could keep pace", might thus have a larger unit consisting of both step-down contours, but each would also be a unit in its own right.

## 4 CONCLUSION

The divergencies discussed in the above sections highlight the theoretical problem underlying the association of auditory and signal-based analysis. At the segmental level, the facility for precise, de-contextualised replay may lead to more closely signal-linked percepts than is possible (or even desirable) in traditional auditory transcription. Also, the demand in labelling to allocate only one symbol to any given stretch of signal contradicts the known parallel nature of information transfer in speech. In fluent continuous speech, this will always lead to conflicts, however clearly segmentation criteria may be formulated. At the prosodic level, it appears that inconsistencies in marking tone group boundaries were mainly a result of the inability of the system as it stands to capture a hierarchical structure. As at the segmental level, overlapping units cannot be consistently forced into a linear string.

## 5 REFERENCES

[1] Beckman, Mary E and Janet B Pierrehumbert, 1986, Intonational Structure in Japanese and English, *Phonology Yearbook*, 3, 255-309.

[2] Roach, Peter J, 1979, *JIPA*, 9,1,2-6

[3] Wells, JC, 1990, *Longman Pronunciation Dictionary*, Longman.

This research was funded by Esprit Project 2589. Symbols are in SAMPA.