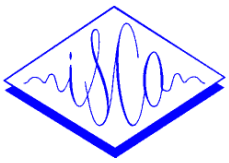

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

RELIABILITY OF SPEECH SEGMENTATION AND LABELLING AT DIFFERENT LEVELS OF TRANSCRIPTION

Barbara Eisen

Institut für Phonetik und Sprachliche Kommunikation
Schellingstr. 3/II, 80799 Munich, Germany

ABSTRACT

The investigation presented in this paper is a first attempt to specify the reliability of manual segmentation and labelling. It is demonstrated that transcription performance does not only rely on external conditions, i. e. segmentation criteria, transcribers' labelling experience or characteristics of labelling devices but rather depends on internal acoustic-phonetic features of the analysed utterance and the level of abstraction of description. Three levels of transcription are introduced. The reliability of segmented data at each level is discussed on the basis of consistency measures within the judgements of different transcribers.

Keywords: *Speech segmentation, labelling, levels of transcription, interindividual consistency,*

10.21437/Eurospeech.1993-163

1. INTRODUCTION

Recent national and multinational activities in creating speech databases gave rise to an increased interest in manually segmented and labelled speech data. As a consequence the practice of transcribing segments has again become subject to critical discussions both from a theoretical and applicational point of view [1].

In addition to several other areas of language research, segmented and labelled speech data have mainly been used for training and testing purposes in ASR systems. To meet the demands of speech technology high reliability of the labelled data has to be achieved. With a special view to the development of standardized labelling procedures an analysis of segmental description was carried out that will give a first outline of the problems a segmental annotation might give rise to with a given labelling procedure.

There are several factors that determine the quality und usability of a segmentation of the speech wave for technical applications, some of which concern mainly external conditions, others that refer to internal aspects. As far as external conditions are concerned, three basic issues have to be addressed. Firstly, a well defined catalogue of segmentation and transcription rules forms the foundation of every segmentation and labelling task. The development of a training session for transcribers on the basis of these pre-defined criteria would then be the next step toward a standardized transcription performance. Finally, the technical equipment used for segmentation has to provide an environment that allows an all-round examination of the speech signal together with an easy handling of all available functions. But even if these basic external conditions are optimal and a transcription performance of high precision can be guaranteed, segmentation and labelling will still be influenced by factors that we have to accept as side-effects of either the degree of abstraction which a description of the physical signal has to achieve or the inherent acoustic-phonetic characteristics of the defined entity (e. g. phone-sized segments, acoustic events) itself. In practice, the degree of abstraction corresponds to what is called "level" of transcription and in general is related to a specific transcription method, as will be seen from the descriptions below.

Three different levels of transcription, the acoustic-phonetic level, the level of narrow phonetic transcription and the phonemic (or broad-phonetic) level will be compared. Since interindividual consistency in transcription and segmentation performance can be regarded as a major cue to reliability, the amount of agreement in re-labelled utterances has been computed. Though transcription and segmentation are closely related they form two distinct aspects of signal description which both bear their own problems. Therefore these tasks will be treated separately.

2. THE LEVEL OF ACOUSTIC-PHONETIC DESCRIPTION

The results of consistency measures that will be presented in this section are based on a set of re-labelled utterances taken from the PHONDAT [3] corpus of read German sentences, all recorded in an acoustically controlled environment. Segmentation at this level required the detection of signal proximal acoustic-phonetic structures called ***events***. The list

of segmented and labelled events that have been selected for this analysis is shown in Table 1. At this stage no explicit claims are made concerning the phonological status of the annotation symbols. Thus one speech sound may be

Table 1: Acoustic-phonetic events

FR	FRICATIVE - source feature *noise*
VD	VOICED - source feature *voice*, presence of periodic excitation
TR	TRANSITIONAL - rapid change in formant frequency
VO	VOWELLIKE - presence of clearly visible formant structure
VQ	VOWEL_QUALITY - clear formant structure with particular formant frequency location

composed of several acoustic events and one single instance of an acoustic-phonetic segment in turn might cover more than one phonemic unit. It should be noted that the transcription strategy at this level of description is special in the sense that all events are represented as a separate tier so that successively, for each of the acoustic properties, transcribers have to pass through the speech signal marking the stretches of speech where they think the particular event

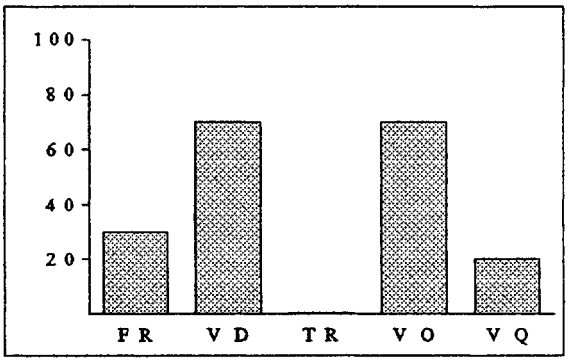

Fig. 1: Percentage of agreeing segmentations concerning the number of detected events in a given utterance

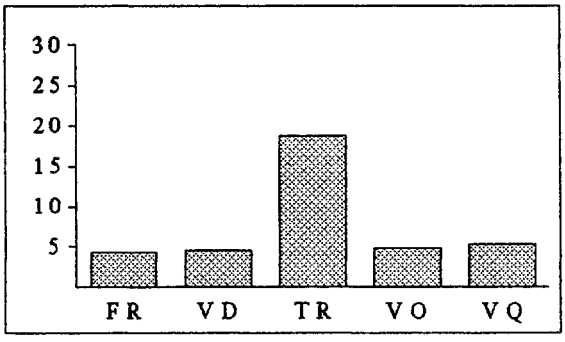

Fig. 2: Mean deviation (in ms) in the placement of initial segment boundaries.

to be present. Hence, the resulting transcript is discontinuous within tiers and overlapping across tiers. Consistency of transcription at this level might be seen in the agreement concerning the number of detected instances of a particular event in a segmented utterance. This type of consistency in acoustic-phonetic labelling is shown in Fig. 1. Obviously, labelling performance differs with respect to the type of acoustic feature to be transcribed. There is relatively high consistency when labelling voiced and vowellike portions of the speech signal, moderate agreement with the events VQ and FR and nearly no agreement if the event is TR. Compared to these results segmentation performance does not show the same kind of dissimilarity over acoustic events involved in this analysis (see Fig. 2). Despite the interindividual differences in the number of marked events, we find a relatively high agreement in the placement of initial segment boundaries. The range between different transcripts is generally below 6 ms. Only the score for the event TR mirrors the weak labelling performance.

3. THE LEVEL OF NARROW-PHONETIC DESCRIPTION

If one proceeds from a pure acoustic description to the segmentation and labelling of phonetic sound categories according to the IPA symbol chart requirements change considerably.

The analysis of labelling performance at this narrow-phonetic description level is based on speech recordings of 100 sentences taken from Sotschek [4], which had been uttered by six speakers in received pronunciation. Each utterance was re-labelled by at least three transcribers independently. Subjects had to give an articulatory-phonetic description of the actual realisation of the words in the sentence and were allowed to assign diacritical marks to the base symbol whenever necessary. Segmentation was continuous und sequential except at word boundaries, where overlapping segments could be defined.

For computation of consistency all symbols of the IPA chart were grouped into a smaller set of phonetic categories (see Table 2) with respect to articulation manner. Only the glottal stop is listed separately. This should account for the fact that in fluent speech this sound is often indicated by a change in voice quality in the neighbouring segmental units. The consistency in transcription was measured by counting the number of completely agreeing annotations in relation to the number of potential occurrences of a given category in the data set. Figure 3 illustrates that there is a strong dependence of interindividual agreement on the phonetic category that has to be transcribed. Speech sounds obviously differ in the extent to which they tend to be labelled identically. Fricatives, nasals and laterals get relatively high scores whereas for vowels, voiced plosives and especially the glottal stop consistency is weak. Taken the consistency measures overall, the agreement between transcribers never crosses the 90% mark. If we compare these findings with the segmentation

Table 2: Phonetic categories

symbol	sound class
Pvl	voiceless plosives
Pvd	voiced plosives
Fvl	voiceless fricatives
Fvd	voiced fricatives
L	laterals
N	nasals
Vf	front vowels
Vc	central vowels
Vb	back vowels
GS	glottal stop

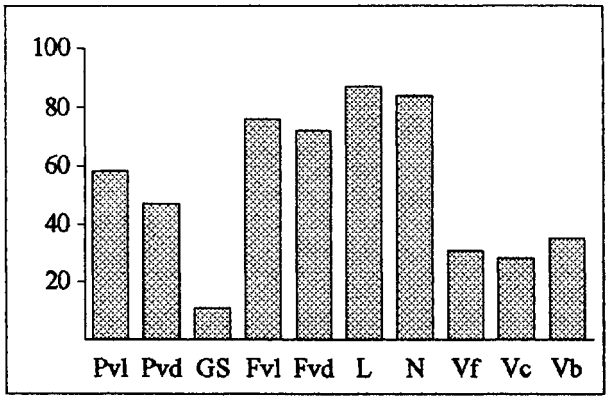

Fig. 3 : Percentage of identical transcriptions with respect to phonetic categories

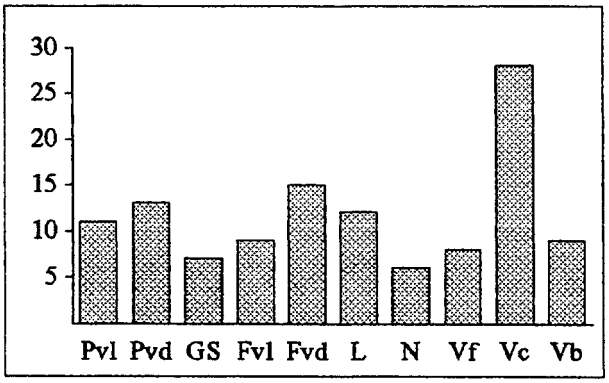

Fig. 4: Mean difference in the definition of initial segment boundary, specified by different phonetic categories. Divergence expressed in ms.

performance (placement of initial segment boundary, as above) we find that high dissimilarity in transcription does not necessarily correspond to low consistency in segmentation (see Fig. 4). The precision with which different transcribers define the initial segment boundaries of speech sounds is reflected in the range of corresponding sample values. Low bars in the graph in other words correspond to high agreement. Consistency measures for the placement of segment boundaries were computed only for those segments that had been transcribed identically. As in the preceding analysis, precision is highly dependent on sound category. For nasals there is an average difference of 6ms, front and back vowels, glottal stops and voiceless fricatives yield values below 10ms, all other categories are below 15ms, except central vowels, where the range is unexpectedly high.

4. BROAD-PHONETIC DESCRIPTION

The third set of re-labelled utterances was based on the same data as described in Section 2. The labelling strategy at this level required a segmentation of the speech wave with respect to the citation form of the words given in a slightly modified SAMPA notation. The elements in this alphabet all have phonemic status, but they could be used for non-phonemic description as well, to denote the actual phonetic realisation of the words. The citation-phonemic labels of a word were shown on the computer screen during a labelling session and transcribers had to decide, whether an offered symbol could be accepted (in which case the actual realisation conformed with the citation form) or whether any kind of deviant pronunciation required the annotation of either a deletion, insertion or sound substitution.

For the computation of consistency values all symbols had again been assigned to a group of sound categories according to Table 2. The agreement in labelling is illustrated in Fig. 5. Overall results show a considerable increase in consistency compared to the narrow-phonetic level. This applies to all sound categories. For fricatives, nasals and laterals more than 90% agreement was achieved and for glottal stops, voiced plosives, and all vowels the gain from this transcription level is exceptionally high. But, as one can see from Fig. 6, this general improvement in transcription does not apply to segmentation performance. The overall results do not exceed the values for narrow-phonetic segmentation. Instead, for voiceless plosives, laterals and the glottal stop a slight increase

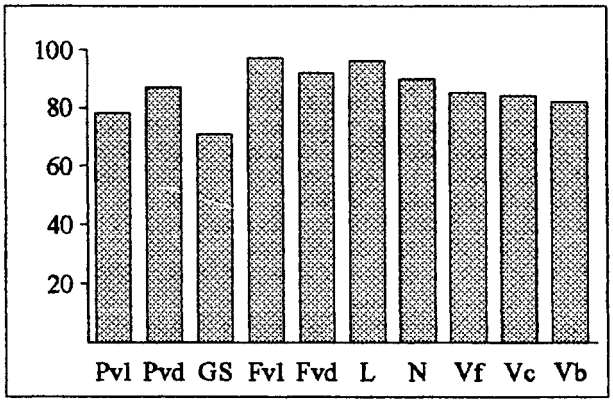

Fig. 5: Percentage of identical transcriptions when broad-phonetic description is performed.

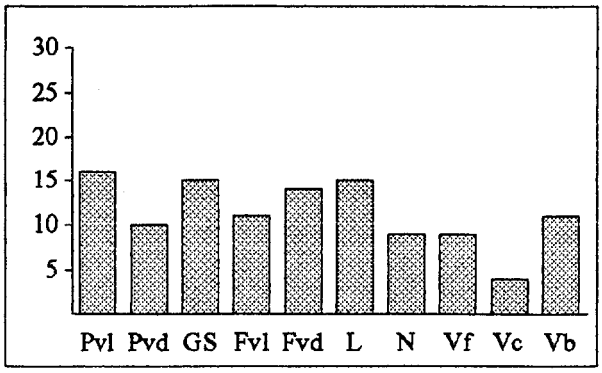

Fig. 6: Interindividual differences in the placement of initial segment boundaries. Average boundary displacement in ms. Equal transcriptions presumed.

in the range of segment initial boundaries can be observed. But a small improvement might be seen in the fact that there is no sound category with boundary displacements of more than 16ms on average.

5. DISCUSSION

Three levels of description have been introduced for a comparison of interindividual consistency in the description of the speech signal. As already argued at the outset, a separate analysis of labelling and segmentation performance is mandatory if one wants to get a better idea of the factors that determine the reliability of segmental transcription. It is obvious from the presented data that segmentation data pattern in a different way from consistency measures of transcription and thus cannot be directly derived. These findings are not surprising if we look at the requirements that go together with either aspect of the description task. Whereas segmentation mainly relies on visual cues of the speech wave or derived signal representations for a correct positioning of segment boundaries, transcription, in the first instance, is performed on the basis of the auditory impression. Since speech sounds differ in their perceptual status (cf. [2], where a distinction is drawn between clear and unclear cases) the dependence of transcription and segmentation on sound category is straightforward. In fact, the relationship between the consistency measures of labelling and segmentation can be either proportional or inversely proportional. As an example for a rather parallel behaviour one might take the measures of the event TR at the acoustic-phonetic level. This acoustic feature, being inhomogeneous by its very nature, is difficult to define both in auditory as well as physical terms, which complicates the detection of the event itself as well as the precise definition of its beginning and end (cf. Fig. 1 and Fig. 2). Laterals, instead, can be easily described in auditory terms (i. e. consistency in transcription is high) but in comparison to transcription, segmentation values signal a relatively low interindividual agreement. A similar relation applies to nasals, when the narrow-phonetic and broad-phonetic description level is compared. From the former considerations it can be inferred, that transcription is much more dependent on the abstraction level, since the degree of abstraction mainly affects the auditory categorisation of speech sounds. Thus a strictly categorical labelling method gains a high amount of consistency.

Though consistency is a desirable aim, one should still be critical of extremely good results. High agreement at higher level representations does not necessarily imply adequacy of transcription. Of course, human listeners may err in assigning labels to the physical signal and consistency in some cases might simply stands for a common error. This presumably applies to some extent to the broad-phonetic transcription strategy described above, where transcribers judgement might be biased by the offered notation symbol which would lead to a higher amount of projective judgements.

6. CONCLUSION

The presented comparison of labelling and segmentation performance makes it clear that there is no general answer to the question of reliability. It revealed that the quality of speech annotations used for technical applications must be seen against the background of description level, inherent perceptual features of the speech sounds in a language, and the requirements of the performed labelling task. The current study is a first step toward a better insight into the problems of segmental description. It was carried out on the assumption that the answers to the question of manual labelling will be a key to an understanding of the underlying mechanism of speech perception in general. The question of whether and how contextual features might affect consistency in transcription and whether the presented results hold for spontaneous speech and a more colloquial speech style remains subject to future investigations.

ACKNOWLEDGEMENTS

My sincere thanks are due to Lioba Faust, Institute of Phonetics, Bonn, for providing the acoustic-phonetic labelling data and to Christoph Draxler at CIS, Munich, who computed the consistency measures by means of Prolog routines.

REFERENCES

[1] Barry, W.J., Fourcin, A.J.: Levels of labelling. Computer Speech and Language 6, 1992, 1-14

[2] Eisen, B., Tillman, H.G., Draxler, Ch.: Consistency of judgements in manual labelling of phonetic segments: the distinction between clear and unclear cases. Proceedings ICSLP 92, Banff, 1992, 871-874

[3] Pompino-Marschall, B. (ed.): PHONDAT. Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesprochenes Deutsch. Forschungsberichte Institut für Phonetik und Sprachliche Kommunikation (FIPKM) 30, 1992, 99-128

[4] Sotschek, J.: Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die deutsche Sprache, In: Fortschritte der Akustik. DAGA 84, Darmstadt, 1984, 873-876