A Multi-Level Automatic Segmentation System: SAPHO and VERIPHONE

C. Dours, M. de Calmès, H. Kabré, J.M. Pécatte, G. Pérennou, N. Vigouroux
IRIT/CERFIA UA 824 CNRS Laboratory
Paul Sabatier University
118 route de Narbonne
31062 TOULOUSE Cedex
FRANCE

ABSTRACT

After having described problems involved in automatic labelling and having shown current application in the domain of automatic speech processing, we expose our automatic labelling system. This system comprises two sub-systems: - the sub-system for labelling and segmenting to phonetic events, - the VERIPHONE system, which aligns strings of phonetic events with strings of phonetic notation.

Results obtained on the EUROM-0 corpus are discussed.

1. INTRODUCTION

Automatic labelling of a speech signal is one of the goals of the ESPRIT project SAM (1) — Multi-lingual Speech Input/Output: Assessment, Methodology and Standardisation — In the framework of acoustic phonetic data base (2), (3), automatic labelling (19), (20), (21), (22) is absolutely necessary: experiences show a considerable gain in time for the labeller when the system proposes labelling flags. Another advantage of automatic labelling is that it always uses the same criteria — stability and reproducibility —. This last argument is of course limited to situations where the system is independent of recording conditions and speaker conditions.

The utilitarian aspects should not however hide the fundamental underlying problems of automatic labelling. In particular, since the problem is one of aligning the phonetic transcription or notation with a phonetic representation of the signal (4), (5), (6), the problems involve the pertinence of phonological and phonetic representations.

The model of the phonological component which makes explicit the relations between these levels of representation brings up some controversial questions:
- about the nature of this type component,
- about its mental reality,
- about intra- and inter-speaker variability
- and about its translation in automatic speech recognition systems (a priori knowledge provided by an expert or knowledge learned through stochastic learning).

This communication aims simply at describing our automatic labelling model and the tools, TEX, SAPHO and VERIPHONE, that we have developed for automatic labelling (cf. Fig. 1). The results obtained on the EUROM-0 corpus seem to be a useful contribution with regards to the questions just mentioned.

2. THE TEX ENVIRONMENT

2.1 Description

TEX is an environment aiding in segmentation, labelling and automatic speech recognition. It allows transforming sequential acoustic input into phonetic units (phonemes, syllables, ...). It comprises a knowledge acquisition and compilation module as well as a transduction module.

Acoustico-phonetic knowledge acquisition and compilation module — It allows the user to define cue representations — static and dynamic acoustic cues — in declarative form, as well as segmentation and labelling strategies in the form of ATNs. Access to TEX's library of pre-defined functions also eases these tasks. For details, one can refer to (7), (8).

Transduction Module — Once the vocal signal, transcribed in the chosen acoustic representation (spectral centiseconds, for example), has been put in the input file, and the knowledge has been compiled, the user can activate the engine, thus producing the corresponding output files (cf. Fig. 2).

Complex applications can involve more than one transduction, operating sequentially and/or in parallel. When the user decides that his cues and segmentation strategy are satisfactory, he can create a "run-time". From here on, we will call SAPHO the run-time that we defined for the phonetic event segmentation of the EUROM-0 corpus.

The TEX environment, written in C language, is available on IBM-PC. A second version for UNIX operating system is being implemented on the workstation MCS600.

EUROSPEECH '89, Paris, France, September 1989 2083
2.2 The SAPHO system: automatic phonetic event segmentation module

The speech signal is digitalized at 16Khz. The acoustic preprocessing consists of a bank of 24 pre-accentuated filters (9). The parameters retained for segmenting are total energy and high frequency energy, both static and dynamic. We also defined two parameters, acuteness and compactness, for identifying the phonetic events.

These parameters were normalized in order to compensate for particular recording conditions and effects related to speaker-dependency. We will see in §4 that a good normalization of the acoustic level minimizes speaker-dependency in the phonetic event segmentation module.

3. THE VERIPHONE SYSTEM

The VERIPHONE system takes as input a string of phonetic events produced by SAPHO, and a phonetic transcription or notation provided by an expert (10), (11). The output is a list of temporal assignations for the phonetic units.

The VERIPHONE system was derived from the VORTEX system, a spelling and typing correction system (12). A detailed description can be found in (13).

VERIPHONE uses three sources of knowledge:
- the source containing the string in phonetic notation X,
- the coder which contains phonetic knowledge expressed as probabilized rewrite rules
- and the canal, which models decoding errors and which contains the confusion matrix as well as costs associated with oversegmentation —case of discrete phonemes— and undersegmentation —case of clusters.

Each phonetic unit is associated with a theoretic tempo, used for the temporal normalization —one can take into account, for example, work published in (14), (15), (16), (17). This normalization process also uses prosodical diacritics, inserted in the phonetic notation, in order to indicate lengthening. As indicated in (5) and (6), it is thus possible to reduce inter- and intraspeaker variability.
In this article, we will only discuss the results obtained on the first sentence of the EUROM-0 corpus for speakers MD, SA and DP.

Structurally, a phoneme can be pronounced as a series of events: this is the case of plosives, vibrated and lengthened sounds, vowels, nasal vowels, etc. Conversely, an event can include more than one phoneme: this is the case for clusters, etc.

The diagrams in figures 4 and 5 illustrate these phenomena for each one of the speakers.
4.2 The VERIPHONE system

In the preceding corpus of the EUROM-0, VERIPHONE grouped the phonetic events into the phonetic units that would have been obtained by manual labelling. The differences between boundaries are on the average 15 ms. Greater differences arise in some difficult cases, for which manual labellers are in disagreement and for which there are no objective cues for determining a boundary.

The phonetic rules that we adopted take into account the statistics given above. Cases of undersegmentation stemming from structural causes are treated by using symbols representing groups of phonemes that tend to be pronounced as a single cluster (ex: (fricative + semivowel), vowel + (liquid coda), ...).

In this brief article, we give an example (see Fig.6) of the labelling table obtained by the VERIPHONE system.

5. CONCLUSION

The system described here allows speaker-independent automatic labelling. These results are made possible through spectral and temporal normalization.

Although VERIPHONE allows adopting a specific set of rules for each speaker, in practice this is not useful. Indeed, the statistical results for each speaker given in Fig.4 and 5 show a good stability in the segmentation of phonetic units into events - segmentation that VERIPHONE's phonetic rules make explicit.

With regards to the questions presented in the introduction, this suggests that variability resulting from acoustic properties, both static and dynamic of the phonation system could be processed during the acoustico-phonetic decoding through dynamic normalization. Another source of variability is located in the phonological component. This variability results from the fact that notations for the same utterance can vary according to the speaker or the context.

This study shows that the decoding necessary for automatic segmentation, which can be speaker-independent, can be thought of as being located between two levels. On the more abstract level it can inhibit phonological variability, and on the more concrete level it can inhibit phonation variability.

Other studies, such as the one reported in (18) do not follow exactly this model, however this work is concerned by analytic recognition of continuous speech. It seems to us that the question of speaker variability — versus speaker-independent decoding variability — still remains an open one.

6. BIBLIOGRAPHY


(6) K. Torkola, "Automatic Alignment of Speech with Phonetic Transcriptions in Real Time", in proceeding IEEE-ICASSP, 1985, pp. 142-146.
(22) H.G. Goldrein, D.R. Reddy, "Feature extraction, segmentation and labelling in the HARPY and HEARSY-II Systems, JASA N° 60.