



SUPERVISED SEGMENTATION WITH APPLICATION TO SPEECH RECOGNITION

X.L. AUBERT *

ABSTRACT

A system is described which performs time-alignment of continuous speech with phonetic transcription. The approach combines several techniques popular in A.S.R. (Dynamic Programming, Clustering) together with the explicit use of speech specific knowledge. The system is speaker independent, fully automatic and is able to cope with phonological variations like elision or assimilation of phonemes and insertion of pause or noise-like segments. It has been tested on several speakers and has proven to be well suited for the direct estimation of parameters required by a statistically-based recognition algorithm, working on a speaker-dependent mode.

INTRODUCTION

Time-alignment of speech signal with phonetic transcription is a well-known problem that has been tackled many times in the past using various methods. A first group of algorithms resorts to time-constrained clustering techniques applied to the whole utterance with little use of the phonetic transcription (ref 1-2, a.o.), while other approaches mainly rely on acoustic-phonetic and phonological knowledge, formalised and structured into a (possibly very complex) set of rules (ref 3-4, a.o.). All methods have to cope with the inherent speech variabilities, one particular aspect being here that what has been effectively pronounced may well differ from the standard phonetic transcription. Other important aspects include the choice of speech units (phonemes, diphones, ...) and the possible dependance with respect to the speaker's voice.

The present approach combines explicit use of speech specific knowledge at various levels within the frame of classical tools like D.P. and has been designed to automatically segment continuous speech in phoneme-like units. The system works on a speaker-independent mode (with no manual extraction of reference prototypes) and is able to adapt -to a certain extent- the standard phonetic transcription to what has been really uttered.

To assert the quality of the segmentation, the system has been used to directly estimate the statistical parameters required by a speaker-dependent speech recognition system based on Hidden Markov Models (HMM, ref 6) of phonemes. As an alternative to iterative training (by Viterbi or Baum-Welch algorithm), the training phase is decoupled in two separate parts namely, the phonetic segmentation of the training sentences and the parameters estimation (by collecting all the samples corresponding to the same phonemic class). For a connected-digit database of five German speakers, this procedure has produced recognition scores which compare favourably with those produced by the standard iterative training.

SEGMENTATION SYSTEM OVERVIEW

Having chosen the phoneme as our basic speech-unit, a set of about 50 phonemes is needed to cover the whole German language. This phoneme set is further organised into Broad Phonetic Classes (BPC, ref 5), which are roughly

* Philips Research Laboratory, Brussels, Av. Van Becelaere, 2-Box 8, B-1170 Brussels - BELGIUM

based on the manner of articulation. Each BPC includes several phonemes sharing gross spectral characteristics in common. Here follows the list of BPC considered in this study; 1: weak events (pause,occlusion); 2: intervocalic liquid and glide; 3: semi-vowels; 4: unvoiced plosives; 5: fricatives (including the voiced fricative 'Z'); 6: voiced plosives; 7: nasals; 8: Front vowels; 9: central vowels; 10: diphthongs; 11: back-vowels; 12: postvocalic liquid and glide.

The segmentation of a given sentence is performed through four consecutive stages. First, a "standard" phonetic transcription is produced from the known sequence of words making up the utterance. Second, the speech signal is aligned with this phonemic sequence by globally optimising an objective function taking spectral similarity, duration distribution and phonological rules into account. Third, starting from this time-alignment, each consecutive phoneme couple is segmented using specialised procedures devoted to particular pairs of BPC. This second pass leads to a refined segmentation while making checks concerning the validity of assimilation or elision rules and boundary locations. Four, a "realised" phonetic transcription is produced together with the corresponding landmarks in the signal and ,possibly, messages in case of problems detected during the process.

STANDARD PHONETIC TRANSCRIPTION

A pragmatic approach has been adopted which consists in concatenating the phonetic transcription of each word -as given by a phonetician- together with the introduction of some special allophones. Although the system is aimed at connected speech, we always insert an "inter-word" symbol to absorb possible breath noise or glottal stop and which may be skipped as well. Plosives are split in their occlusion and burst portions, the realisation of each one being controlled by context-dependent rules allowing for incomplete closure or unreleased burst. Besides, distinction is made between the intervocalic and postvocalic 'R' and 'L', the latter being treated as an extension of the preceding vowel.

GLOBALLY OPTIMISED TIME-ALIGNMENT

Each phoneme of the language is characterised by 3 (single) values: a loudness index, a spectral-distribution index and a typical duration. These parameters provide enough information for the time-alignment task and make unnecessary the extraction of reference prototypes.

Let i denote the i -th "centi-second frame" in the signal, j the j -th symbol in the transcription and $p(j)$ the corresponding phoneme. Then a spectral similarity measure between the i -th frame and the j -th symbol is given by:

$$d[i,j] = C1[p(j)] | LR[p(j)]-LF[i] | + C2[p(j)] | SDR[p(j)]-SDF[i] | ,$$

where $C1[.]$, $C2[.]$ are weights depending on the phoneme $p(j)$,

$LR[.]$, $SDR[.]$ are the reference values provided by a table for (resp.) the loudness and spectral distribution parameters,
 $LF[.]$, $SDF[.]$ are the current values of the same parameters calculated at a particular centi-second frame in the signal,
 $| . |$ denotes the absolute value operator.

The spectral distribution parameter is obtained from the first normalised auto-correlation coefficient while the loudness parameter comes from the root mean square energy, both parameters being processed through adaptive clipping and scaling to provide a kind of voice normalisation.

Concerning duration modelling, each phoneme $p(j)$ is characterised by a reference duration value $DR[p(j)]$ provided by the same table as before. The first

step consists in adapting these values to the sentence averaged speaking rate: Let ρ denote the ratio between the expected length (obtained by summing the DR[.] for all the symbols appearing in the transcription) and the observed length of the sentence (obtained through endpoint detection on the signal). Then, for all phonemes but a few exception (plosives), the duration values are linearly adapted following: $\alpha[p(j)] = DR[p(j)] / \rho$.

The second step consists in associating to each $\alpha[.]$ a duration distribution; we use discrete Poisson distribution (ref 6) which are modified for some phonemes (fricative, long vowels) to cope with strong lengthening effects.

Now, using $d[i,j]$ as a local distance and the negative logarithm of the duration distribution as an additive weight, the globally optimised time-alignment can be obtained by a dynamic programming algorithm (ref 6, a.o.). The phonological rules are embedded in the D.P. recurrence to allow for the skipping of certain phonemes: interword symbol, first word phoneme, occlusion, burst.

LOCAL SEGMENTATION PROCEDURE

For each pair of consecutive phonemes, an accurate boundary location is performed by means of specialised BPC-procedures which are applied on the respective sub-interval determined by the global time-alignment. It is thus assumed that the latter is accurate enough to insure that the true boundary lies within this sub-interval. A particular procedure is devoted to each of the following cases that cover most of the possible phonemic contexts: Voiced-Unvoiced pair, Occlusion-Burst sequence, Burst-Fricative sequence, Vowel-Nasal pair, Vowel-Liquid pair, Vowel-Glide pair and Diphtong segmentation.

For each procedure, the general principle is the following (ref 4) : within the pre-assigned sub-interval, specialised features are extracted and processed into an appropriate similarity measure. A time-constrained clustering algorithm provides the boundary location which is checked with respect to a minimum and maximum allowable duration. In case of failure, a message is produced and an ad hoc solution is applied (for exemple, a vowel-liquid sequence is divided in two equally long parts).

SEGMENTATION RESULTS

The system has been tested on a connected digit database where the 10 vocabulary words are transcribed using a subset of 22 phonemes, and on a phonetically-balanced database including 341 vocabulary words and all phonemes of the German language. For the 6 speakers (4 males, 2 females) treated so far, no significative differences consecutive to their voice details or speaking rate were observed in the segmentation results. On a total of 500 sentences with an average duration of 4 seconds, it appears that the coarticulation rules are appropriately used by the system to adapt the phonetic transcription. Due to the duration control of each phoneme, the system is free of gross segmentation errors; however, there are still small problems in situations like a nasal followed by a liquid or a glide, a sequence of 2 or 3 plosives and for the postvocalic 'R' or 'L'. The corresponding local procedures are currently being improved. Although the program structure is fairly developed, its execution is very fast: about 3 times real-time on a Vax-780.

APPLICATION TO CONNECTED SPEECH RECOGNITION

Hidden Markov Models of phonemes are widely used for connected speech recognition. In most of the cases, the statistical parameters have to be trained on the particular speaker's voice and this asks for a complex iterative procedure. This led us to investigate the feasibility of estimating directly the required parameters from the automatic phonetic segmentation without any

additional step, thus providing a kind of a posteriori performance test. The experiments have been performed on a seven-digit phone number database, each of the 5 speakers (3 male, 2 female) having fluently spoken 100 such utterances. Speech is described every centi-second by the 15 first cepstral coefficients, the gain term being discarded.

Four cases were considered, depending on the type of training procedure and the number of states per phoneme:

- HMM (standard case, ref 6) with Viterbi training and 3 states per phoneme;
- SMM (Semi-Markov Models with state-occupancy modelling, ref 6) with Viterbi training and 3 states per phoneme;
- SEGMENTATION with direct parameter estimation and 3 states per phoneme, each segment being divided in 3 equal parts;
- SEGMENTATION with direct estimation and 1 state per phoneme, repeated 3 times in the word models;

Results are summarised in the following table (D=deletion, I=insertion, S=substitution):

RECOGNITION RESULTS: Nber of ERRORS at WORD LEVEL for 100 seven-digit STRINGS				
SPEAKER	HMM-3 states	SMM-3 states	SEGMENT.-3 states	SEGMENT.-1 state
M1	3 (OD, 2I, 1S)	1 (OD, 1I, 0S)	0 (OD, 0I, 0S)	0 (OD, 0I, 0S)
M2	7 (OD, 0I, 7S)	1 (OD, 0I, 1S)	2 (1D, 0I, 1S)	6 (2D, 1I, 3S)
M3	1 (OD, 0I, 1S)	1 (OD, 1I, 0S)	0 (OD, 0I, 0S)	0 (OD, 0I, 0S)
F1	9 (OD, 6I, 3S)	4 (OD, 0I, 4S)	1 (OD, 0I, 1S)	3 (OD, 2I, 1S)
F2	23(1D, 11I, 11S)	UNAVAILABLE	12(1D, 4I, 7S)	13(OD, 9I, 4S)

The best results are obtained by the direct estimation training; even the "1 state per segment" model outperforms the HMM-3 states.

CONCLUSION

The main observations to be drawn from this study are: the usefulness of elementary phonological rules for adapting the phonetic transcription and the appropriateness of a simple spectral similarity measure when combined with phonemic duration modelling. A priori segmentation followed by direct parameter estimation seems a viable alternative worthwhile to be pursued on more complex recognition tasks.

ACKNOWLEDGMENT

This work was sponsored by the Belgian Ministry of Economic Affairs under IRSIA-IWONL grant N° 4819. Only the author is responsible for the content of this publication. The author would like to thank his colleague Herve Bourlard for supplying him with the results produced by the HMM recognition system.

REFERENCES

1. J.S.Bridle and N.C.Sedgwick, ICASSP 1977, p 656.
2. K.Maenobu, Y.Ariki and T.SAKAI, Information Sciences 33, 1984, p 31.
3. R.De Mori, P.Laface and Y.Mong, IEEE Trans. on Pat. An. Mach. Intell. Vol 7-1, 1985, p 56.
4. H.C.Leung, "A procedure for automatic alignment of phonetic transcriptions with continuous speech", S.M. thesis, Mass. Inst. Technol., Cambridge, MA, 1985.
5. V.W.Zue, Proceedings of the IEEE, VOL. 73-11, 1985, p 1602.
6. H.Bourlard and C.J.Wellekens, "Connected Speech Recognition By Phonemic Semi-Markov Chains For State Occupancy Modelling", EUSIPCO-86, Third European Signal Processing Conference, The Hague, The Netherlands, 1986, p 511.