ON THE USE OF PHONE DURATION AND SEGMENTAL PROCESSING TO LABEL SPEECH SIGNAL

Philippe Depambour, Régine André-Obrecht*, Bernard Delyon***
depambou@irit.fr, obrecht@irit.fr, delyon@irisa.fr
*IRIT - UMR CNRS 5505, 118 route de Narbonne, F-31062 Toulouse Cedex, FRANCE
** IRISA, Campus universitaire de Beaulieu, F-35042 Rennes Cedex, FRANCE

ABSTRACT

This paper presents recent work on continuous speech labelling. We propose an original automatic labelling system where elementary phone models take a segmental analysis and the phone duration into account. These models are initialized by a short speaker-independent training stage in order to constitute a model database. From the standard phonetic transcription, phonological rules are gathered to process the various pronunciations. For each new corpus or speaker, a new quick unsupervised adaptation stage is performed to re-estimate the models, and then follows the correct labelling. We assess this system by labelling a difficult corpus (sequences of connected spelled letter) and sentences of one speaker of the BREF80 corpus. These results are quite promising, in the two experiments less than 9% of phonetic boundaries are incorrectly located.

1. INTRODUCTION

Many areas in speech research, like speech analysis, speech synthesis and speech recognition, require large segmented and labelled databases. Phonetic labelling of speech is usually performed manually by phonetic experts. This procedure is extremely time consuming and the size of these databases makes a manual labelling difficult. So an automatic procedure for transcribing speech seems to be the best way to avoid these inconvenients. When the procedure is efficient, the phonetic expert has only to verify the results and correct the wrong marked boundaries or bad labels.

During the last decade, many systems have been proposed; they are based on Hidden Markov Models [1], Neural Network Models (NN) [2] or Knowledge-Based system [3]. Recently, others methods like Context Sensitive Rules [4] or even Dynamic Time Warping [5] have been studied to improve efficiency. The task remains difficult; especially most of those systems require an important training stage, and a first hand labelling.

The labelling system we develop in this paper uses the Markov modelling: a segmental HMM takes the speech signal quasi-stationary property and the phone duration into account. Furthermore, the training stage is replaced by an automatic adaptation to a new corpus and a new speaker without performance degradation.

2. SYSTEM OVERVIEW

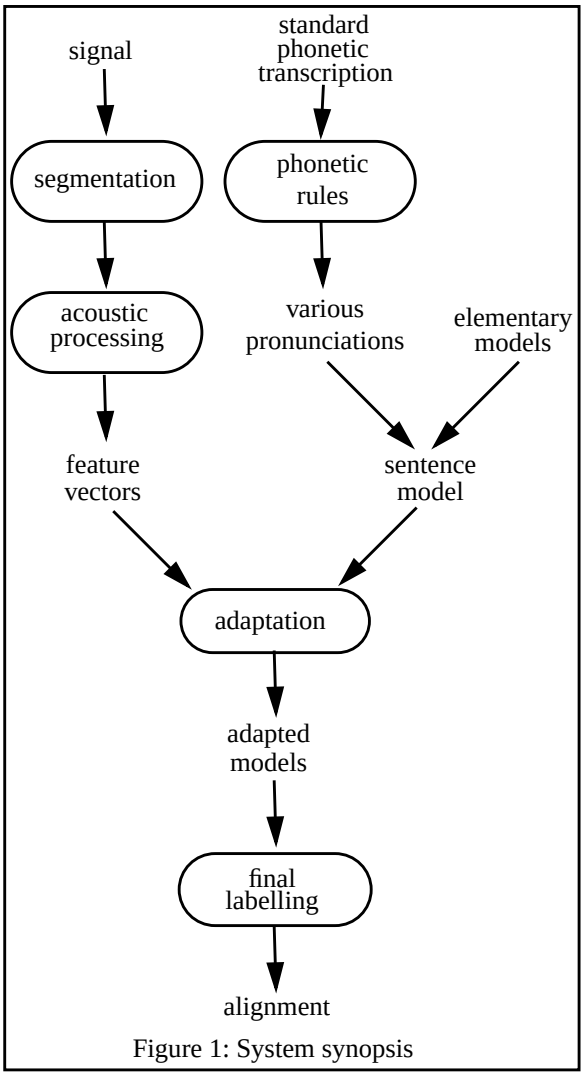


Figure 1: System synopsis

As any labelling system, the input is the speech signal and its high-level phonetic transcription; the output is the matching sequence of speech segments with the associated low-level phonetic transcription.

The labelling procedure involves in three steps:
- an automatic *a priori* statistical segmentation is performed on the signal. Then, for each segment, an observation vector is computed;
- a global statistical model is built from the phonetic transcription, a set of phonetic rules and elementary phonetic models using phone duration;
- the alignment is obtained via a modified Viterbi procedure.

The Figure 1 shows the description of our system.

2.1 Acoustic Segmental Preprocessing

The speech signal is assumed to be described by a string of homogeneous units, thus an automatic *a priori* segmentation is performed on the signal using the «Forward-Backward Divergence» algorithm [6]. This method gives us a succession of stationary segments; they correspond to the steady part of a phone (when it exists) and transient segments. Consequently, a phone is described by only two or three segments.

The preprocessing consists in an analysis of each segment where Mel frequency cepstral coefficients are extracted. The segment energy and the segment duration are added to create a 9-coefficient vector. This observation vector can be divided into a cepstral part $Y \in \Re^8$ and a prosodic part: the variable segment duration $T \in \Re$.

2.2 HMM and Phone Duration

The Hidden Markov Models (HMM) are successfully performed in speech recognition system. But a major deficiency of a classic HMM is that both spectral and prosodic features are uniformly processed. In particular they can not appropriety model the temporal structure of speech. Therefore the global phone duration is an important cue to discriminate between phonetic units since this feature is correlated with the phone identity.

We propose here an alternative to the Two-Level HMM described in [7] which take the global phone duration into account. Each phone ϕ is modeled by:

- a classical HMM where $p^{\phi}(t)$ is the probability of a transition t, and $p_t^{\phi}(Y)$ is the probability of observing Y on the transition t;
- $d_{\phi}(\tau)$, a probability density function to model the duration τ of the phone ϕ. It is not a state duration but the global phone duration.

We note these phone models M_{ϕ}^{d}. The likelihood of the observations sequence $O_{1\ldots N} = ((Y_1,T_1),\ldots,(Y_N,T_N))$ is given by

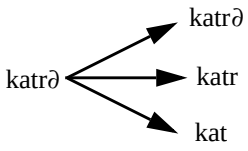
$$P(O_{1\ldots N}) = \sum_{(t_1\ldots t_N)} \left(\prod_{i=1}^{N} p^{\phi}(t_i) p_{t_i}^{\phi}(Y_i) \cdot \prod_{j=1}^{P} d_{\phi_j}\left(\sum_{k \in A_j} T_k \right) \right)$$

where $(\phi_1,\phi_2,\ldots,\phi_P)$ is the phone sequence associated to the transition sequence $(t_1,t_2,\ldots,t_N)$, and where A_j is the set of consecutive indices i such that the transition t_i belongs to the phone ϕ_j.

Thanks to the segmentation, each phone consists of a maximum of 3 or 4 parts (except the silence), so we can build the M_{ϕ}^{d} models without loop and each phone transition is tied to the number of observations emitted in the same phone.

The global statistical model of any new sentence is obtained by applying phonological rules to the standard transcription to form the most common possible pronunciations and by concatening the corresponding elementary models M_{ϕ}^{d}.

For instance, the digit "4" may have three different pronunciations in french:



To find the best alignment, an usual Viterbi algorithm can be used by adding up an obligatory transition at the end of each path inside a phone model with d_{ϕ} as probability density function (see Figure 2).

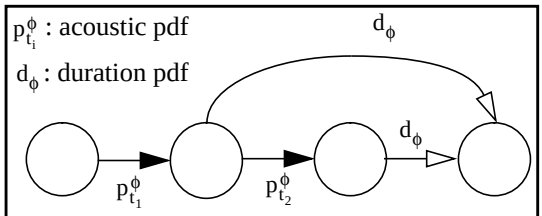


Figure 2: An elementary phone model

Every path through such model is a succession of acoustic emissions which is ended by a duration emission.
So, if we note s the final state of a transition t, and

$$P(s,n) = \max_{t_1\ldots t_n}^{t_n = t} \left(\prod_{i=1}^{n} p^{\phi}(t_i) p_{t_i}^{\phi}(Y_i) \cdot \prod_{j=1}^{q} d_{\phi_j}\left(\sum_{k \in A_j} T_k \right) \right)$$

where the phone sequence $(\phi_1,\phi_2,\ldots,\phi_q)$ was completely

visited through the path $(t_1,t_2,\ldots,t_n)$, the Viterbi algorithm becomes:

- if s is strictly within a model

$$P(s,n) = \max_t(P(s',n-1)\cdot p^{\phi}(t)p_t^{\phi}(Y_n))$$

- if s is a final state

$$P(s,n) = \max_t\left(P(s',n-1)\cdot p^{\phi}(t)d_{\phi}\left(\sum_{i=n-j+1}^{n} T_i\right)\right)$$

where j is the number of transitions visited in ϕ .

The acoustic probability density functions are Gaussian distributions. But the results shown in [8] lead us to choose as duration probability functions inverse Gaussian distributions.

3. ADAPTATION PROCEDURE

A short speaker-independent training stage has been performed to initialize the set of elementary phone models to obtain a preliminary database of models which distributions are speaker and application independent.
These training data have consisted of 58 CVCV nonsense sequences, 11 sentences containing nasal vowels, 80 isolated short words containing a sequence stop-R or R-stop, pronounced by a female speaker and 30 phonetically well-balanced sentences pronounced by the same female speaker and a male speaker. The procedure used is a classical EM algorithm where the Viterbi algorithm is replaced by our modified Viterbi algorithm.

The labelling of a new corpus pronounced by a new speaker is performed in two steps:

- the elementary models are adapted to the new speaker by using our training iterative procedure on the whole corpus; only two or three iterations are necessary to obtain a good adaptation;
- a last alignment performed with the adapted models gives the final labelling.

This procedure, entirely unsupervised, takes advantageously the place of an expensive standard training stage.

4. EXPERIMENTS

In order to evaluate the labelling system we have achieved two experiments: the first one with the AMIBE corpus and the second one, actually in progress, with the Bref80 corpus.

4.1 AMIBE [9]

This corpus, recorded at 16 Khz by a male speaker, is composed of sentences of 4 connected letters spelled in French. The labelling has been made of 206 sentences, that is to say 824 letters (about 1700 phones). To present results, we have randomly extracted 48 sentences (about 200 phones) and manually labelled them. Our assessment is based on the delay between the hand-made boundary and the automatic one for the same label; we observe deletions and omissions. We comment these results on the section 4.3. The Table 1 shows the error rate (ER), deletion and insertion rates.

Corpus	ER (> 20 ms)	ER (> 30 ms)	Del.	Ins.
AMIBE	8 %	6 %	3 %	1.2 %

Table 1: AMIBE results

4.2 BREF80 [10]

BREF80 is a large read-speech corpus for French. 80 speakers read texts from French newspaper «Le Monde» without specifying the punctuation. We consider as a new corpus the set of 43 sentences pronounced by the speaker i1f. Each sentence is transcribed in terms of phone by a grapheme-to-phoneme system [11]. Comparisons between automatic labelling and manual labelling have been made on 7 sentences (3 of them contain more than 150 phonemes). You can see the results on Table 2 and a labelling instance on Figure 3.

corpus	ER (> 20 ms)	ER (> 30 ms)	Del.	Ins.
BREF80	8.6 %	6.5 %	1.9 %	1.2 %

Table 2 : BREF80 results

4.3 Discussion

In the two experiments, our results prove the efficiency of our approach. The error rates are entirely satisfactory.
However, we extract some insertions or deletions. They are due to an alignment on an incorrect pronunciation selected among the whole set of pronunciations created by the use of the phonological rules. For instance, a /r/, strongly co-articulated as in the sequence /rl/, could be omitted. It's the same phenomenon with the sequence /tr/. This problem comes from the fact that an expert can locate the phone on the signal thanks to the adjacent phones since it has an influence on them; the automatic system's task is more difficult: how set boundaries when there is no static presence? In fact, an automatic labelling system will always have this sort of restraints. Moreover,

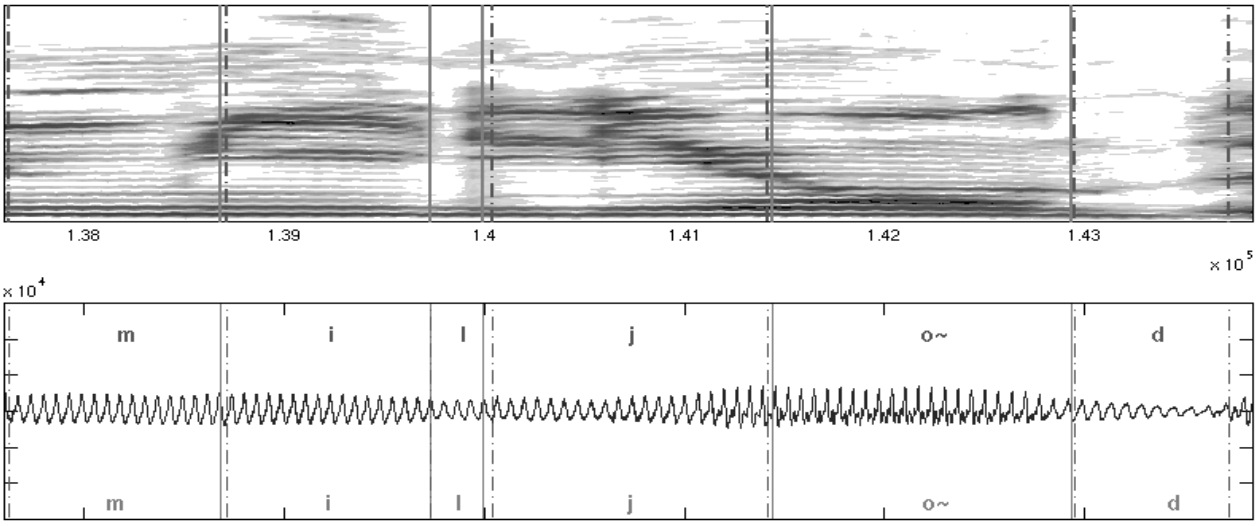


Figure 3: A labelling instance on the French word sequence "million de"
- manual labelling: dashdot line and higher labels
- automatic labelling: solid line and lower labels

in speech recognition or speech synthesis, a specific unit can be used to model those sequences with efficiency and the problem is not so drastic.

5. CONCLUSION

We have presented a new model to take the phone duration into account in a HMM and its utilization, associated with a segmental processing, in a speaker-independent and application-independent automatic labelling system. The obtained results are very promising and our system's performances allow the labelling of large continuous speech corpora.
Our aim now is to improve the phonological rules and the topology of our elementary models. The integration of the models M_{ϕ}^{d} in a recognition task will follow.

6. REFERENCES

[1] D. Fohr, J-F. Mari, and J-P. Haton, "*Utilisation d'un MMC pour l'étiquetage automatique et la reconnaissance de BREF80*", Proc. XXes Journées d'Étude sur la Parole, pp. 339-342, Avignon, France, June, 1996.
[2] P. Dalsgaard, O. Andersen, and W. Barry, "*Multi-Lingual Label Alignment Using Acoustic-Phonetic Features Derived by Neural-Network Technique*", Proc. ICASSP'91, Vol. 1, pp. 197-200, May, 1991.
[3] A. ghio and M. Rossi, "*A Knowledge-based Model for Speaker-independent Acoustic-Phonetic Decoding*", Proc. EUROSPEECH'95, Vol. 1, pp. 807-809, Madrid, Spain, September, 1995.
[4] O. Oppizzi, D. Fournier, Ph. Gilles and H. Méloni, "*Décodage acoustico-phonétique flou*", Proc. XXes Journées d'Étude sur la Parole, pp. 293-296, Avignon, France, June, 1996.
[5] P. Di Cristo and D. Hirst, "*Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable*", Proc. 4ième Congrés Français d'Acoustique, Vol. 1, pp. 425-428, Marseille, France, April, 1997.
[6] R. André-Obrecht, "*A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals*", IEEE Transactions on Acoustics, Speech, Signal Processing, Vol. 36, pp. 29-40, January 1988.
[7] N. Suaudeau, "*Un modèle probabiliste pour intégrer la dimension temporelle dans un système de reconnaissance automatique de la parole*", Thèse de doctorat, Université de Rennes 1, Rennes, France, March, 1994.
[8] N. Suaudeau and R. André-Obrecht, "*An Efficient Combination of Acoustic and Supra-Segmental Informations in a Speech Recognition System*", Proc. ICASSP'94, Vol. 1, pp. 65-68, Adelaide, South Australia, April, 1994.
[9] B. Jacob and R. André-Obrecht, "*Direct Identification vs Correlated Models to Process Acoustic and Articulatory Informations in Automatic Speech recognition*", Proc. ICASSP'97, Vol. 2, pp 999-1002, Munich, Germany, April 1997.
[10] L. F. Lamel, J-L. Gauvain and M. Eskénazi, "*BREF, a Large Vocabulary Spoken Corpus for French*", Proc. EUROSPEECH'91, Vol. 2, pp 505-508, Genova, Italy, September, 1991.
[11] B. Prouts, "*Contribution à la synthèse de la parole à partir du texte : transcription graphême-phonème en temps réel sur microprocesseur*", Thèse de doctorat, Université de Paris XI, Paris, 1980.