IMPROVING SPEECH RECOGNITION WITH MULTIMODAL ARTICULATORY ACOUSTIC HMMs

B. Jacob - C. Sénac - R. André-Obrecht - F. Pellegrino
IRIT- Université Paul Sabatier - CNRS URA 1399
118,route de Narbonne
31062 Toulouse Cedex FRANCE

ABSTRACT

This paper describes a new scheme for robust speech recognition systems where visual information and acoustic features are merged. A segmental processing and two decoding strategies based on Hidden Markov Models (HMM), are studied and evaluated on connected word recognition applications.

INTRODUCTION

The proposed recognition system is one of the components of the AMIBE project (Applications Multimodales pour Interfaces et Bornes Evoluées). The purpose of this work, supported by the PRC's Informatique (Coordinated Research Programs of the CNRS) is to study the natural visual and auditive bi-modality of oral communication and to propose more robust speaker verification and speech recognition systems.

It's well known that, listening in adverse acoustic environments (noise, multiple speakers...) relies heavily on the visual input to disambiguate among acoustically confusable speech elements [1]. To take this phenomena into account, we develop an Automatic Speech Recognition system which processes the synchronization of the 'labial reading' and an acoustic pattern recognition system using HMM.

The lip-reading consists in a pre-processing of the visual information, thus producing a set of articulatory features as described in [2]. The acoustic pre-processing is based on a segmentation algorithm followed by a cepstral analysis. But as articulatory target positions and acoustic steady segments are not always synchronized, we propose two different strategies for merging these two kinds of data:

- a concatenation of the cepstral and labial vectors which provides a global observation vector for a classical HMM;
- a master/slave type relationship between two HMMs [3] which leads to correlate the two informations.

RECOGNIZER OVERVIEW

An automatic speech recognition system involves basically two components : the prepocessing to reduce the information and the linguistic decoder.

Extraction of the signal parameters

Our recognition system processes two kinds of signal :
- an acoustic signal sampled at 16 kHz,
- three articulatory signals composed of the lip breadth (A), the lip height (B) and the lip area (S), sampled at 50Hz [4].

The acoustic signal is preprocessed by an automatic segmentation [5] and a spectral analysis is performed on each segment. Therefore 8 Mel frequency cepstral coefficients (MFCC) are extracted. We add the energy (E) and their first derivatives (8 δMFCC, δE).

The acoustic segment boundaries are projected on the articulatory signals; for each segment, we calculate the mean of each labial parameter and their first derivatives.

The global feature vector consists of 18 acoustic coefficients, 6 articulatory ones, and the duration of the segment (T). Figure 1 gives an example of an acoustic signal preprocessed by the automatic segmentation

Statistic models of the linguistic decoder

Two different approaches are proposed :

- a global standard H.M.M., M_{glob}, is hierarchically built ; each word model is obtained by concatenation of elementary acoustic models. The elementary unit is the 'pseudo-diphone' ; it corresponds to the steady part of a phone or the transient parts between adjacent sounds and the acoustic model is a basic left to right continuous density H.M.M. ;
- in the master/slave approach, two parallel H.M.Ms are built. The first one, named articulatory H.M.M. M_{art}, is an ergodic model of three states and three pdfs, which takes the articulatory features into account. The second one, named acoustic H.M.M. M_{acous}, has the same topology as M_{glob} and processes the acoustic observations only. The M_{art} H.M.M. controls the M_{acous} HMM, in the sense that the M_{acous} H.M.M. transition and observation probabilities depend on the current state in M_{art} [3].

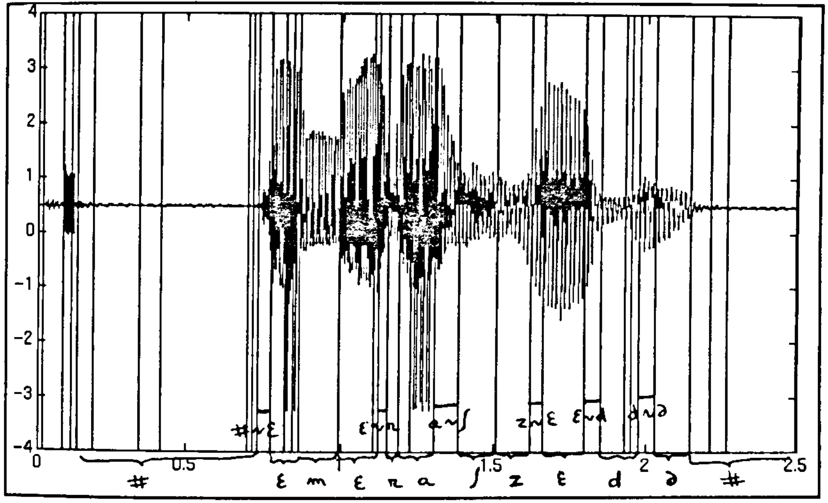

Figure 1: Example of a preprocessed acoustic signal by automatic segmentation. The four spelled letters "M R H Z" are pronounced. For each segment, is indicated the pseudo diphone label found by the Viterbi algorithm, with the M_{glob} model :
# represents 'silence';
# ~ɛ , ɛ ~ m, ɛ ~ r, a ~ʃ ... represent transient units,
a, d ... represent steady units.

EXPERIMENTATIONS

We have experimented these two approaches on two mono speaker applications : connected digit recognition and connected spelled letter recognition. The connected digit corpus is composed of sequences of four digits : 84 sentences for the learning set and 35 sentences for the test set. The connected spelled letter database is composed of sequences of four spelled letters : 158 sequences for the learning set and 48 for the test set.

Table 1 gives the error numbers on the digit test set, in terms of sentences and words. It is well known that very good results are obtained with a classical HMM such as M_{glob} and with 8 standard MFCCs. We observe no performance loss when using the segmental processing (1 word substitution) and a very small one when introducing the labial information (3 word substitutions). The comparison between the classical HMM M_{glob} and the master/slave M_{acous} +M_{art}, shows a better result for the global approach (3 errors vs 5 errors), but this remark must be qualify : first the confidence interval doesn't permit a precise conclusion, and then the complexity of the master/slave HMM is such that the number of parameters is too important to hope a good learning with such a little learning set

Table1: Recognition error number, in terms of sentences and words, in accordance with the parameters and the models.

Model	Coefficients	Sentences /35	Words /125
M_{glob}	8MFCC+E+T	1	1
	8MFCC+E+T +A+B+S	3	3
master/slave $M_{acous} + M_{art}$	8MFCC+E+T +A+B+S	5	5

For the spelled letter application, an initial M_{glob} model is learned with 8 MFCC, the energy and the segment duration ; we add successively the 3 labial parameters and their derivatives. The same experiment is repeated with an initial global HMM learned with 8 MFCC, the four first derivatives, the energy and its derivative, and the segment duration. The results are reported on Figure 2. We observe that the best recognition rate is obtained when using the lip height and breadth. The introduction of the lip area doesn't bring any pertinent information, it is strongly correlated with the parameters A and B. The derivatives appear as noisy information, the desynchronization between the labial information and the acoustic one is certainly one of the cause. On Figure 1, we can see the alignment of the sentence "M R H Z" obtained by the Viterbi algorithm through the best M_{glob} model (8MFCC, E, T, A, B), in terms of pseudo-diphone units ; segments and pseudo diphone units are perfectly aligned.

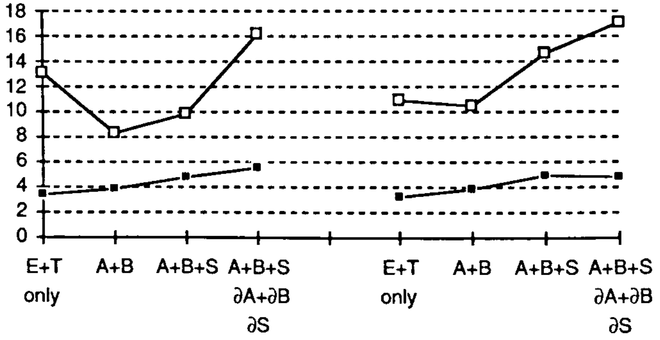

Figure 2: Recognition with the M_{glob} HMM : Word error rate for the learning set (■) and for the test set (□) in accordance with the articulatory vector. On the left part of the figure, the acoustic vector consists of 8 MFCC, E, T as on the right part, the first four derivatives are added.

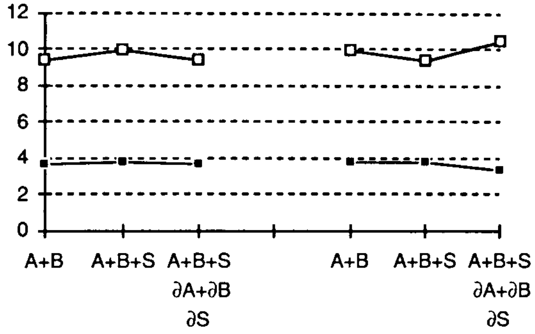

Figure 3: Master/slave M_{acous} +M_{art} HMM results : word error rate for the learning set(■) and for the test set (□) in accordance of with the articulatory vector. On the left part of the figure, the acoustic vector consists of 8 MFCC, E, T as on the right part, the first four derivatives are added.

The same experimental protocol is performed to test the master/slave approach ; of course, we add to the initial parameters the labial ones, A and B. The results are reported on the Figure 3. We notice that the results are quite so good as using the global approach and they are more steady : the introduction of supplementary parameters (labial or acoustic derivatives) doesn't disturb the recognition rate , in view of the confidence interval. This remark is very promising, we may hope that this structure is the best one to introduce the labial information and that it will be more robust when the acoustic parameters will be noisy. Future experiments must confirm this conclusion.

CONCLUSION

We have described two statistical approaches based on HMMs to merge articulatory and acoustic information and to improve an automatic speech recognition. Experimental results show the difficulty to process the desynchronization between the lip moves and the acoustic signal. It seems that the more robust approach is the master/slave one, future studies must confirm this hypothesis.

REFERENCES

[1] P.Duchnowski, Meier U., Waibel A. : "See me, hear me: integrating automatic speech recognition and lip-reading". S11-6.1 ICSLP 94, YOKOHAMA.

[2] C. Benoit , Abry C., Boe L.J. : "The effect of context on labiality in french". Eurospeech 91, GENOVA.

[3] F. Brugnara, de Mori R., Guiliani D., Omologo M. : "A family of parallel HMM". p 1103, Proc.IEEE Int. Conf. ASSP 92, SAN FRANCISCO.

[4] M.T. Lallouache: "Un poste "visage parole" couleur. Acquisition et traitement automatique des contours de lèvres". Thèse de Doctorat de l'Institut National Polytechnique de Grenoble, 1991.

[5] R. André-Obrecht: "A new statistical approach for the automatic segmentation of continuous speech signals". IEEE Trans. on Acoustics, Speech, Signal Processing, vol 36 n°1, January 1988.