DIRECT IDENTIFICATION VS CORRELATED MODELS TO PROCESS ACOUSTIC AND ARTICULATORY INFORMATIONS IN AUTOMATIC SPEECH RECOGNITION

Régine André-Obrecht[1] *Bruno Jacob*[1]

[1]IRIT-University Paul Sabatier-CNRS UMR 5505, 118 route de Narbonne, F-31062-Toulouse, France

ABSTRACT

Our work deals with the classical problem of merging heterogenous and asynchronous parameters. It's well known that lips reading improves the speech recognition score, specially in noise condition ; so we study more precisely the modeling of acoustic and labial parameters to propose two Automatic Speech Recognition Systems :

- a Direct Identification is performed by using a classical HMM approach : no correlation between visual and acoustic parameters is assumed.

- two correlated models : a master HMM and a slave HMM, process respectively the labial observations and the acoustic ones.

To assess each approach, we use a segmental pre-processing. Our task is the recognition of spelled french letters, in clear and noisy (coktail party) environments. Whatever the approach and condition, the introduction of labial features improves the performances, but the difference between the two models isn't enough sufficient to provide any priority.

1. INTRODUCTION

It's well known that lip reading improves the human speech recognition performance, crucially in noise conditions. Consequently, the multimodal aspect of the speech perception has been widely studied, specially these bimodal, optic and acoustic, stimuli, and the corresponding visual and auditory systems [9], [5] during the last years.

More recently have appeared Automatic Speech Recognition system which integrate acoustic and visual speech signals. The approaches are classical : Artificial Neural Networks [2] , Hidden Markov Models are the most currently used. In the last category, we find the following systems :

- Direct Identification (DI) Model where only one HMM is used and its input observation vectors are the simple concatenation of the acoustic and visual vectors, considered as independent [1].
- Two independent HMMs processing separately the data flows ; then a decision rule is applied on each score [7].
- An HMM product is built from a visual HMM and an acoustic HMM ; the data are processed simultaneously [8]

These approaches don't take the anticipation and retention phenomena between the phonatory organs into account, except in the last case where it is trained automatically. The conclusions remain shy. Our work deals with another way of combining the labial and acoustic informations, and handling the asynchrony. We propose a segmental analysis to process both acoustic and labial informations. Then we study two linguistic decoders : the first one is the classical DI model and the second one is based on two correlated parallel HMMs, in order to exploit viseme units and acoustic pseudo-diphone units asynchronously.

We have assessed and compared our systems using a connected spelled french letter recognition task. Many experiments have been performed in clean and noisy environments (cocktail party noise), to fix up the recognizers.

2. TWO SEGMENTAL MODELS : DI AND MASTER-SLAVE MODELS

As we say previously, to merge acoustic and labial features, we suggest and compare two systems. Each one involves basically two components, a segmental pre-processing and a statistical linguistic decoder, for which we study a Direct Identification Model and a Master-Slave Model.

• 2.1. The signal pre-processings

The pre-processing is shared by the two proposed recognizers. The acoustic signal is automatically segmented by the Forward-Backward divergence method [10], without *a priori* knowledge. A sequence of acoustic steady and transient segments are obtained (Figure 1). A 16ms window is centered on each segment and a cepstral analysis is performed to provide 8 MFCC and the energy E. A regression upon the adjacent windows gives the derivatives of these parameters (8 ΔMFCC, ΔE).

The visual input consists of three parameters carefully extracted from a front view of the speaker lips [6]. They correspond to the three main characteristics of lip gestures [3], namely internal lip width A and height B, and intero-labial lip area S. They are stored every 20 ms along the speech waveform. The boundaries given by the acoustic segmentation are projected on the labial

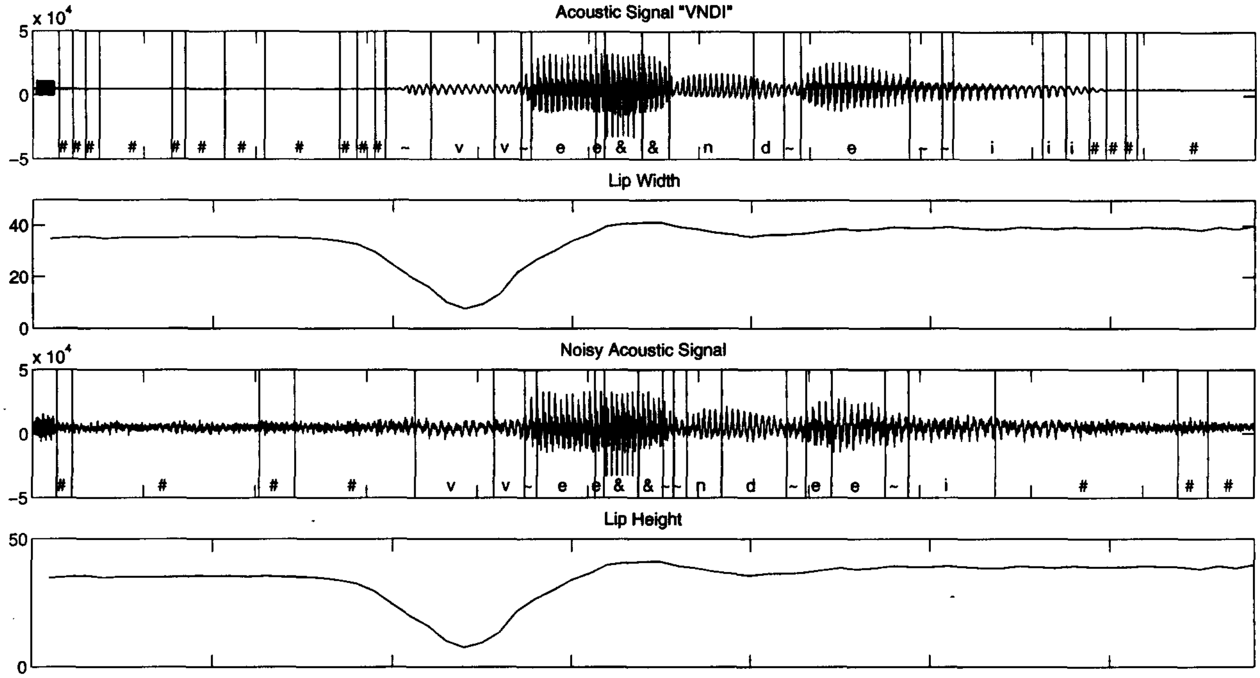

Figure 1. ***Segmental pre-processing of the acoustic signal and the noisy signal (cocktail party SNR= 10dB), the lip width and the lip height curve of the sentence "VNDI" /ve ndei/. The trajectories found by the MS model Viterbi algorithm are mentionned in terms of pseudo diphones units, (a) in clean conditions (b) in noisy environnements (10dB).***

	8MFCC	8MFCC+T	8MFCC+T+E	8MFCC +A+B	8MFCC+T+E +A+B	8MFCC+T+E+4ΔMFCC +A+B
clean conditions	88%	93%	93%	95%	96%	96%
noise SNR=10dB	80%	86%	87%	86%	88%	91%

Table 1. ***Recognition Rates using the Direct Identification Model.***

signals, and for each segment, means and derivatives ($\hat{A}$,$\hat{B}$,$\hat{S}$, ΔA,ΔB, ΔS) are computed.

Finally, the pre-processing module provides to the decoder, for each segment, an input vector of 25 components, corresponding to 18 acoustic coefficients, 6 labial ones, at which the segment duration (T in ms) is added. The pre-processing is used during the training phase and the recognition phase.

• 2.2. The Master Slave HMM

The statistical model of the linguistic decoder is based on two correlated parallel HMMs (Figure 2):

- the Master HMM is a classical HMM of three states and three pdfs, which correspond to characteristic visemes : open (o), semi-open (so) and close (c) lips. The observation vector is composed of the 6 labial coefficients per segment ($\hat{A}$,$\hat{B}$,$\hat{S}$, ΔA,ΔB, ΔS).
- the Slave HMM is built hierarchically by introducing as elementary units *the pseudo-diphones*, ie. the steady parts of phone or the transitions between two phones. Each unit is modeled by a very simple HMM (1 pdf per model) and transient ones may be omitted. Each word of the application is described with these units, taking variable pronounciations and coarticulations into account. The slave observation vector consists of the acoustic vector and the segment duration (8 MFCC, E, 8 ΔMFCC, ΔE, T).

The originality of this approach is that the parameters (transition matrix and pdfs) of the acoustic HMM are **probabilistic functions of the states of the master model.** The theorical properties may be found

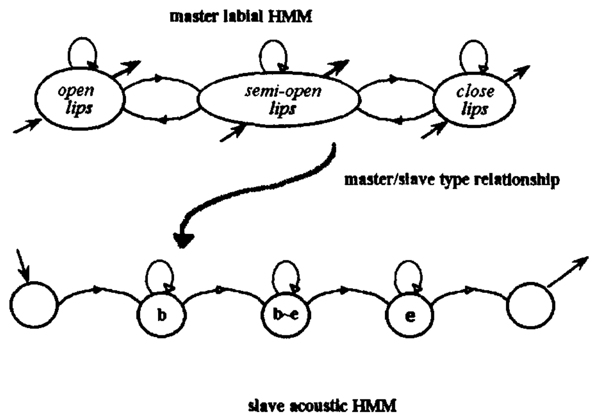

Figure 2. ***A exemple of Master-Slave HMM corresponding to the modeling of the word "B" = /b//e/. Each pdf and each transition probability of the slave HMM depend of the parameter X of the Master HMM, $X \in \{so, o, c\}$***

	8MFCC +A+B	8MFCC+T+ +A+B	8MFCC+T+E +A+B	8MFCC+T+E+4ΔMFCC +A+B
clean conditions	92%	93%	93%	95%
noise SNR=10dB	90%	88%	88%	88%

Table 2. ***Recognition Rates using the Master Slave Model.***

in [4].

- **2.3. The Direct Identification Model**

  The DI Model is a classical HMM. We retain the same topology as the Slave Model one. The observation vector is the concatenation of the labial and acoustic ones, which are assumed independent.

For each approach, pdfs are simple gaussian laws with diagonal covariance matrix.

3. EXPERIMENTS AND RESULTS

Our experimental task is the recognition of the 26 french spelled letters. The sentences are sequences of four connected letters. The training database is composed of 158 sentences (632 letters) and the test one of 48 sentences (192 letters); the experiments are mono speaker.

Many experiments are performed to find the best configuration of the Master-Slave recognizer : it appears that three pdfs are sufficient to characterize the labial information, an increase of the pdf number or a more complex topology don't improve significantly the performances. (All the Master Slave Model results reported in this paper are issued from a labial model of three states).

An other sequence of experiments are performed to find the best family of acoustic and labial parameters:

- When we use the parameter S (alone or with A and B), no improvement is observed; this fact corroborates the correlation between these parameters (S=kAB, where k is speaker dependent).

- We note no significant performance increase when we add all the derivative parameters (acoustic and labial) though they are very useful in classical centisecond HMM.

To confirm the relevance of the visual information in noisy environment, we have added a "cocktail party" noise to the acoustic signal with a 10 dB SNR. The segmentation results are quite robust as the recognition rates. The more significant results are reported on tables 1 and 2.

The difference between the Direct Identification and the Master Slave approaches isn't significant, in view of the confident interval, it's difficult to validate one of them. In clean conditions, the best recognition rate is 96% for the DI model and 95% for the MS model, with the same input vectors (8MFCC, T, E, 4ΔMFCC, A, B). In noisy environnement, the best recognition rate is 90% for the two approaches, the inputs are different: 8MFCC, T, E, 4ΔMFCC, A, B for the DI model and 8MFCC, A, B for the MS model.

We mustn't forget that the number of parameters to be learned is more important for the MS model than for the DI model, and our database is too limited to correctly learn much more ones. It may be an explanation for the small recognition rate difference between the two models, and the relative stability of the MS model performances when the input vector dimension increases. Other database are presently recorded, and future experiments will lead us to conclude.

4. CONCLUSION

In every configuration (acoustic parameters, environments), integrating the visual information improves greatly the recognition performances. It's a promising research area to obtain more robust recognition systems. So we continue our study with the collaboration of phonetician experts to increase our knowledge about the correlation between visual and acoustic features, and to improve our statical models. New corpora will be also studied.

REFERENCES

[1] Adjouani A. and Benoit C. Audio-visual speech recognition compared accross two architectures. *EUROSPECH 95*, pages 1563–1566, September 1995. Madrid.

[2] Yuhas B.P. and Sejnowski T.J. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, pages 65–71, 1989.

[3] Abry C. and Boe L.J. Laws for lips. *Speech Communication*, 5:97–104, 1986.

[4] Brugnara F., De Mori R., Guiliana D., and Omologo M. A family of parallel hidden markov models. *ICASSP 92*, 1992. San Francisco.

[5] Robert-Ribes J., Schwartz J.L., and Escudier P. A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, pages 1–23, 1994.

[6] Lallouache M.T. *Un poste "Visage-parole" couleur. Acquisition et traitement automatique du contour des lèvres.* PhD thesis, INPG, 1991.

[7] Deleglise P., Rogozan A., and Alissali M. Asynchronous integration of audio and visual sources in bimodal automatic speech recognition. *EUSIPCO 96*, September 1996. Trieste.

[8] Jourlin P., El-Bèze M., and Méloni H. Integrating visual and acoustic informations in a speech recognition system based on hmm. *ICPhS 95*, 4:288–291, August 1995. Stockholm.

[9] Summerfield Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell, editors, *Hearing by eye: the psychology of lipreading*, 1987.

[10] André-Obrecht R. A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(1):29–40, January 1988.

