Words on Lips : How to Merge Acoustic and Articulatory Informations to Automatic Speech Recognition

Régine André-Obrecht, Bruno Jacob, Christine Sénac
IRIT- CNRS UMR 5055 - Université Paul Sabatier
118, route de Narbonne, 31062-Toulouse CEDEX, France
e-mail : obrecht@irit.fr

ABSTRACT

Our work deals with the classical problem of merging heterogeneous and asynchronous parameters. It's well known that lip reading improves the speech recognition score, specially in noisy conditions ; so we study more precisely the modeling of acoustic and articulatory parameters to propose new Automatic Speech Recognition systems. We use a segmental pre-processing, a robust unit "the pseudo-diphone" and we compare a global HMM and a master-slave HMM. We confirm through experiments the importance of labial features in clean and noisy environment.

1 Introduction

It is largely accepted that speech communication is multimodal : the principle "stimuli" are oral and visual "stimuli". Cognitive research has shown that the auditive recognition is greatly improved when the auditor sees the lips of the speaker, and that it is more benefit when he sees the whole figure, specially in noisy environment [Sumby 54], [Benoit 94].

Classical Automatic Speech Recognition (ASR) systems are now effective, but their performances decrease in adverse conditions. The purpose of recent researches is to extract acoustic parameters more robust to the noise or to adapt models to every new condition [Siohan 95].

An alternative approach consists of introducing new parameters as the characteristics of the lips. But this raises two main questions, "How do humans fuse the auditory and visual informations ? At what level ?" and " How can ASR systems integrate them ? Has an automatic system to copy the human behavior ?"

Studies in perception have tried to answer the first question and suggest various integration models [Robert-Ribes 94].

Previous works have shown the ability of Artificial Neural Networks [Yuhas 89] [Watanabe 90], and Hidden Markov Models (HMM) to integrate acoustic and visual speech signals into an ASR system and to increase the recognition scores specially in noisy conditions. When HMM are used, three alternatives have been studied :

- a Direct Identification Model. One HMM is performed and the observations are composed of the concatenation of visual and acoustic data, the fusion process takes place before the classification [Adjoudani 95].
- a Separated Identification Model. An acoustic HMM and a visual HMM process separately and respectively the acoustic data and the visual ones, then a decision is applied (expert rules, combination of probabilities or fuzzy scores) [Foucault 96].

  These two approaches are automatic versions of the models proposed by perception studies.
- a Product Model. A global HMM is obtained as the product of two HMMs (product measure), the observation inputs are composed of the concatenation of the visual and acoustic data and the decoding is global through the HMM product [Jourlin 95].

In spite of these tentatives, the difficulties to merge automatically such hetereogenous data remain unsolved : the signal sampling rates are very different, visual observation extraction is issued from a sophisticated processing [Lallouache 90], and the phenomena of lips retention and anticipation, is not actually correctly modeled [Abry 86].

Our work deals with another way of combining these two streams of information. We suggest a speech recognition system composed of a segmental pre processing of the two signals, and a linguistic decoding based on a master-slave HMM [Brugnara 92] and a pseudo-diphone representation of words : our purpose is to correlate the two identifications during the identification process.

In this paper, after a description of the signal preprocessings and a presentation of the decoder, we comment some results when the application is the automatic recognition of connected spelled French letters. We compare this approach to the classical one where a global observation vector (concatenation of the labial vector and the acoustic one) is decoded by a global centisecond HMM, for two different environments.

These recognition systems have been studied within the framework of the AMIBE project (Applications Multi-

modales pour Interfaces et Bornes Evoluées) supported by the PRC's Informatique (Coordinated Research Programs of the CNRS).

2 Description of the ASR systems

As we say previously, to merge labial and acoustic features, we suggest and compare four systems. Each one involves basically two components : a pre-processing and a linguistic decoder.

2.1 The signal pre-processings

In what concerns the visual input, we have three labial signals (one triplet each 20ms) which correspond to three main characteristics of lip gestures [Abry 86], namely horizontal width (A) and vertical height (B) of the internal lip opening, and internal lip area (S). Note that these parameters are correlated as future experiments show it. The acoustic signal is sampled to 16kHz.

We have distinguished and compared two pre-processings :

- *the classical centisecond analysis.* For each acoustic frame of 16 ms, 24 channel Mel power spectrum is obtained by applying triangular windows on the FFT output. From this, 8 Mel Frequency scale Cepstrum Coefficients (MFCC) are computed. A regression upon five adjacent frames gives their derivatives (Δ MFCC). At these 16 coefficients are appended the energy of the frame and its derivative (E, Δ E). Parallelly, an interpolation computed upon each labial signal provides every 16ms three features A, B, C and a regression gives ΔA, ΔB, ΔS.

- *the segmental analysis.* The acoustic signal is automatically segmented by the Forward-Backward divergence method [André-Obrecht 88], without *a priori* knowledge. A sequence of acoustic steady and transient segments are obtained (figure 1).

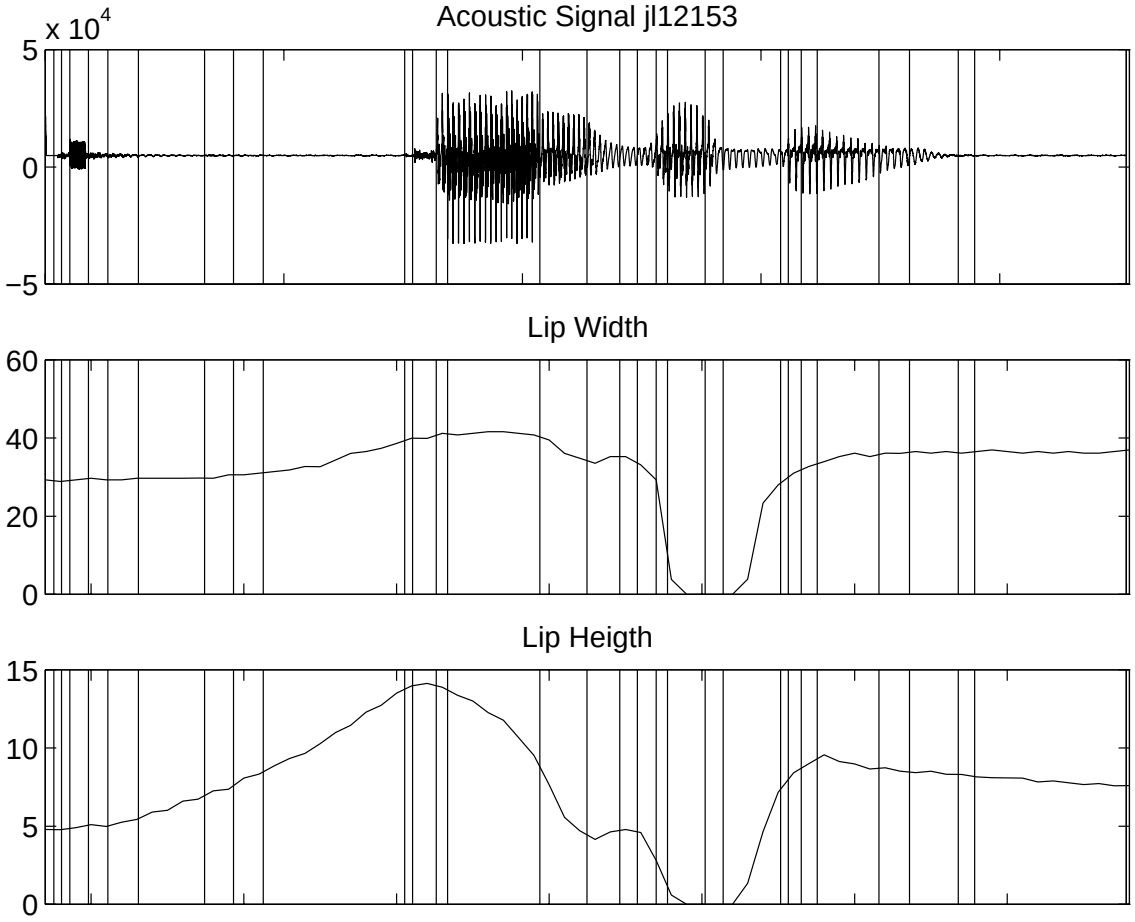

Figure 1: Segmental pre-processing of the acoustic signal and the labial curves

A 16ms window is centered on each segment and the previous 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 boundaries of the acoustic segmentation are projected on the labial signals, and on each segment, means and derivatives (ΔA, ΔB, ΔS) are computed.

Finally, the pre-processing module provides a vector of 24 components, corresponding to 18 acoustic coefficients, 6 labial ones. For the centisecond approach, it is the feature vector of a frame of 16ms, for the segmental one, it's this of a segment. In this last case, the segment duration (T in ms) is appended to the vector. **The same pre-processing is used during the training phase and the recognition phase**.

2.2 The linguistic decoder

The statistical model of the linguistic decoder is based on HMMs. We compare two kinds of models :

- The Reference Model M_{ref}. It is built hierarchically by introducing the elementary unit : *the pseudo-diphone*, a steady part of phone or a transition between two phones. Each word of the application is described with these units. Each unit is modeled by a very simple HMM (1 pdf per model, for the segmental analysis and 3 for the centisecond one) except for silence where we use 2 pdfs. The observation input may be the previous feature vector or only of its components.

- The Master-Slave HMM $M_{art} - M_{acous}$. It is based on two parallel HMMs, a master model which is a classical HMM and a slave model whose parameters (transition matrix and pdfs) are probabilistic functions of the state of the master model [Brugnara 92]. For our purpose, the master model, the articulatory model M_{art}, is an ergodic model of three states and three pdfs. The observation input is composed of the 6 labial coefficients. The slave model, the acoustic model M_{acous}, has the same topology as the Reference Model M_{ref}, but its observation input is composed only of the acoustic parameters (and eventually of the segment duration).

3 Experiments

3.1 Database and assessment of the Reference Model

Our experimental application is the recognition of the 26 french spelled letters. The sentences are sequences of four connected letters and the experiment is mono speaker. The training database is made of 158 sentences (632 letters) and the test one, 48 sentences (192 letters). To assess the Reference Model, we use three different observation inputs : the segmental labial inputs, the segmental acoustic inputs and the centisecond acoustic inputs, we make the dimension of the input vector vary. We report in table 1 the best performances obtained by each category, and the description of the input vector.

Recognition system	Training set	Test set
labial M_{ref}		
A B	51 %	40 %
A B ΔA ΔB	48 %	39 %
segmental M_{ref}		
8MFCC E T	98.4 %	90.1%
centisecond M_{ref}		
8MFCC E	95 %	89.4 %

Table 1: Best Recognition Rates using the Reference Model.

Perception results are confirmed by the first experiment : the lip reading allows about 40 % of good recognition and two parameters as A and B are sufficient. S is too correlated to A and B to bring more informations. The best performance is obtained by a segmental pre-processing, it proves that the pseudo-diphone unit is very appropriate to this pre-processing. With the derivatives (labial and acoustic), we observe a decrease of the recognition rates, the database may be not large enough.

3.2 M_{ref} versus $M_{art} - M_{acous}$

We use the labial and the acoustic information simultaneously, when using the two kinds of pre-processing and the two kinds of linguistic decoders, in clean environments. For every configuration, we make the dimension of the labial input vector and this of the acoustic one vary ; in this paper, we give the best combinations.

Recognition system	Training set	Test set
segmental		
8MFCC, E, T		
A, B	97.2 %	91.8%
centisecond		
8MFCC, E, 4ΔMFCC, ΔE		
A	96.6 %	94.2 %

Table 2: Best Recognition Rates obtained by the reference model M_{ref} in clean conditions

The table 2 shows the recognition rates when using the Reference Model. When comparing them to the scores reported in table 1, we observe an increase of the performances by introducing one or two labial parameters (A, B). This increase is more important with the centisecond pre-processing.

Recognition system	Training set	Test set
segmental 8MFCC, E, T A, B, ΔA, ΔB	99.2 %	91.8%
centisecond 8MFCC, E, 4ΔMFCC, ΔE A, B	98.1 %	96.5 %

Table 3: Best Recognition Rates obtained by the master-slave model in clean conditions

The table 3 shows the same results when using the master slave model ; we observe, for the centisecond pre processing, an increase still more important. We notice the introduction of the cepstral derivatives is necessary only in the case of the centisecond analysis.

3.3 In adverse conditions

We add a "cocktail party" noise to the acoustic signal (SNR = 10dB). The segmentation process is not efficient enough to use in these conditions. So we present here the assessment of the two linguistic decoders combined with the centisecond analysis (table 4).
Obviously we observe a decrease, and the Reference Model seems to give better results. Therefore the difference between the training recognition rate and the test one, leads us to think the database isn't large enough to correctly learn a master-slave model whose parameter number is much bigger. It is a real difficulty to have efficient labial parameters and large database !

Recognition system	Training set	Test set
M_{ref} 8MFCC, E A, B	89.7 %	76 %
$M_{art} - M_{acous}$ 8MFCC, E, 4ΔMFCC, ΔE A, B, ΔA, ΔB	95 %	71 %

Table 4: Best Recognition Rates obtained in noisy conditions.

For all the experiments, the derivatives don't bring any great improvement, but this may be also explained by the previous reasons, a too small database. Other experiments will be performed to justify this assertion.

4 Comments

We have described several statistical approaches to merge articulatory and acoustic informations. We have compared them to a classical centisecond acoustic approach based on HMM. We prove that the best performances are obtained when introducing a labial vector.
The scores we observe for the segmental analysis are quite good, but we hope much more : a previous study shows that the segmental approach may provide quite good results, specially with telephonic speech, and we have proved the pseudo-diphone unit is quite adequate to this pre-processing. The cocktail noise is the first case where the segmentation is not so efficient, so the first next experiments must clarify this point.
The second point concerns the master-slave model ; its scores are good but they appear a little worse than the Reference Model ones : as previously say, the number of parameters of the slave model model is very important, the database is not sufficient to learn them. Currently a new database has been recorded, the conditions are more realistic. New experiments will be performed and other noisy conditions will be experimented.

References

[Abry 86] C. Abry, L.J. Boë, Laws for lips, *Speech Communication*, 5, pp 97-104, 1986.

[André-Obrecht 88] 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^o1, janvier 1988.

[Benoit 94] C. Benoit, T. Mohammadi, C. Abry, Audio-visual intelligibility of French speech in noise, *Journal of Speech Hearing Research*, 1994.

[Adjoudani 95] A. Adjoudani, C. Benoit, Audio-Visual Speech Recognition Compared across two architectures, *Eurospeech 95*, Madrid, September 1995, pp 1563-1566.

[Brugnara 92] F. Brugnara, R. De Mori, D. Guiliana, M. Omologo, A Family of Parallel Hidden Markov Models, *ICASSP 92*, San Francisco, 1992.

[Foucault 96] A. Foucault, P. Deléglise, Système acoustico-labial de reconnaissance automatique de la parole, *XXIes JEP*, Avignon 1996.

[Jourlin 95] P. Jourlin, Automatic bimodal speech recognition, *ICPhS95*, Stockholm, august 1995.

[Sumby 54] W.H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America*, 26, pp 212-215, 1954.

[Robert-Ribes 94] J. Robert-Ribes, J.L. Schwartz, P. Escudier, A comparison of Models for Fusion of the Auditory and Visual Sensors in Speech Perception, *Artificial Intelligence Review*, pp 1-23, 1994.

[Watanabe 90] T. Watanabe, M. Khoda, Lip-reading of Japanese Vowels Using Neural Networks. *Proc. ICSLP 90*, Kobe, Japan, pp 1373-1376, 1990.

[Yuhas 89] B.P. Yuhas, M.H. Goldstein, T.J. Sejnowski, Integration of Acoustic and Visual Speech Signals Using Neural Networks, *IEEE Communications Magazine*, pp 65-71, 1989.