HIDDEN MARKOV MODELS MERGING ACOUSTIC AND ARTICULATORY INFORMATION TO AUTOMATIC SPEECH RECOGNITION

Bruno JACOB, Christine SENAC

IRIT- CNRS UMR 5055 - Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse CEDEX France
e-mail : senac@irit.fr

ABSTRACT

This paper describes a new scheme for robust speech recognition systems where visual information and acoustic features are merged. Using as robust unit the « pseudo-diphone », we compare a global Hidden Markov Model (HMM) and a Master/Slave HMM through a centisecond preprocessing and through a segmental one. We confirm by experimentation the importance of articulatory features in clean and noisy environments.

1. 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 bimodality of oral communication and to propose more robust speaker verification and speech recognition system.

It's well known that listening in adverse acoustic environments (noise, multiple speakers ...) relies heavily on the visual input to disambiguate among acoustically confused speech elements [1]. So, how can Automatic Speech Recognition Systems integrate the oral and visual information ?

Previous works have shown the ability of Artificial Neural Networks [2] or of Hidden Markov Models [3] to merge heterogeneous parameters. However most models use either a direct identification from concatenated auditory and visual inputs or a separate identification followed by further combination. Though, these approaches don't permit to solve difficulties to merge heterogeneous parameters and particularly asynchronous parameters due to the phenomenon of lips retention and anticipation.

So we propose an Automatic Speech Recognition System (ASRS) composed of a preprocessing and a linguistic decoder whose approach is based on two HMM built in parallel we named Master/Slave HMM [4].

Experimental results are presented after a concise description of the signals preprocessing and of the linguistic decoder.

2. RECOGNITION OVERVIEW

As many ASRS, ours involves two components : the preprocessing to reduce the information and the linguistic decoder.

2.1. The Signals Preprocessing

As detailed in [5], our recognition system processes two kinds of signals :

- concerning the visual input sampled at 50Hz, we have three articulatory features ; the lip breadth (A), the lip height (B) and the lip area (S) corresponding to main characteristics of lip gestures [6].

- concerning the acoustic signal sampled at 16KHz, we have developed two preprocessings. The first one consists in a classical centisecond analysis where 8 Mel frequency cepstral coefficients (MFCC) are extracted. We added the energy (E) and their first derivatives (8 Δ MFCC, ΔE). The acoustic features are projected on the articulatory signal, for each centisecond are calculated the three labial parameters and their first derivatives. The second one consists in a segmental analysis where the acoustic signal is automatically segmented [7]. As for the centisecond approach, 18 acoustics coefficients and 6 articulatory ones are produced to whose is appended the segment duration (T in ms).

The same preprocessing is applied during the training and the recognition phases.

2.2. The Statistic Models of the Linguistic Decoder

Our 'HMM compiler' [8] permits us to build two kinds of model , so we used two approaches:

- Used in a classical context application, the HMM compiler provides a global standard HMM, named Mglob which 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 to the transient part between adjacent sounds and the acoustic model is a basic left to right continuous density HMM ;

- in the Master/Slave approach, two parallel HMMs are built with the help of the HMM compiler. The first one, the articulatory HMM named Mart, is an ergodic model of three states taking into account the articulatory features. The second one, the acoustic HMM named Macous, has the same topology as Mglob and processes the acoustic observations only. The Mart HMM controls the Macous in the sense that the Macous HMM's transition and observation probabilities depend on the current state in Mart.

3. EXPERIMENTAL RESULTS

3.1. Database

Our application is the recognition of the 26 french spelled letters. The sentences are composed of four connected letters and the experimentation is monospeaker.

Two monospeaker corpus where used :

- the first one named 'Corpus1' was recorded in clean conditions in a specialized laboratory with sophisticated materials. The training set is composed of 158 sentences (632 letters) and the test set of 48 sentences (192 letters).

- the second one named 'Corpus2' was recorded by another speaker with the help of a cam-recorder which introduced noise not only for the acoustic signal but also for the visual one. Here the training set is composed of 130 sentences and the test one of 70 sentences.

3.2 Experiments with Corpus1

3.2.1. Evaluation of the model

Many experiments were made, the first one consisting in evaluating the Mglob HMM with only articulatory inputs and only acoustics inputs. Best performances obtained by each input and the description of the input vector are illustrated in table1. We find again the perception results: lip-reading only allowing about 40% of good recognition. We can add that lip area (S) doesn't bring any pertinent information because it is strongly correlated with the parameters A and B. Using the derivatives, we have observed a stagnation indeed a decrease of the recognition rates.

Recognition System	Training set	Test set
Articulatory Mglob (centisecond) A B	51%	40%
Acoustic Mglob (segmental) 8MFCC E T	98.4%	90.1%
Acoustic Mglob (centisecond) 8MFCC E	95%	89.4%

Table1 : Best recognition rates for the Mglob HMM with only one kind of input

3.2.2. Recognition in clean conditions

Here both articulatory and acoustic information where used with the two kinds of preprocessing and the two kinds of linguistic decoder which give four combinations whose best results are illustrated in table2 .

Recognition System	Training set	Test set
Centisecond Mglob 8MFCC E 4 ΔMFCC ΔE A	96.6%	94.2%
Segmental Mglob 8MFCC E T A B	97. 2%	91.8%
Centisecond Master/Slave 8MFCC E 4 ΔMFCC ΔE A B	98.1%	96.5%
Segmental Master/Slave 8MFCC E T A B ΔA ΔB	99.2%	91.8%

Table2 : Best recognition rates for the Mglob HMM and for the Master/Slave HMM in clean conditions

Comparing the results obtained in table2 with those reported in table1, we note an increase of good recognition when introducing the labial parameters A and B. Here, the centisecond approach seams to produce best results than the segmental one for both the global and the Master/Slave models.

3.2.3. Recognition in adverse conditions

Another experiment consisted in adding a « cocktail party » noise to the acoustic signal to obtain a SNR of 10db. The preprocessing used was the centisecond approach while the two linguistic decoders were used. We can see on table3 the best recognition rates obtained by these two decoders.

Recognition System	Training set	Test set
Acoustic Mglob (centisecond) 8MFCC E	65 %	43%
Centisecond Mglob 8MFCC E A B	89.7%	76%
Centisecond Master/Slave 8MFCC E 4 ΔMFCC ΔE A B ΔA ΔB	95%	71%

Table3 : Best recognition rates in adverse conditions

The first result of the table corresponds to the Mglob processing only acoustical features. The rate of recognition is only of 43%. When adding the articulatory parameters we can see an important increase of the recognition on lines 2 and 3. However we can

observe an important disparity between the recognition rate of the test set (only 76% and 71%) and the training set whose results are not so damaged (89.7% and 95%) : this phenomenon could be explained by a too small database preventing a correct learning.

3.3. Experiments with Corpus2

This corpus was recorded in more realistic conditions than the first one. So, when listening and looking at the acoustic signal, we can observe the same events as for the Corpus1 when adding noise. We can see on table4 the different results of the experiments.

Recognition System	Training set	Test set
Acoustic Mglob (centisecond) 8MFCC E	81.79 %	45.76%
Centisecond Mglob 8MFCC E 4 ΔMFCC ΔE A B	94.7%	74%
Segmental Mglob 8MFCC E T 4 ΔMFCC ΔE A	95.89%	61.86%
Centisecond Master/Slave 8MFCC E 4 ΔMFCC ΔE A B	93.5%	54.7%
Segmental Master/Slave 8MFCC E T A B	95%	61.2%

Table4 : Best recognition rates for Corpus2

As for the Corpus1 when adding noise, we can also deduce that articulatory features bring an important increase of the recognition rate. Here as well, we can observe a big disparity between results obtained for the test set and for the learning set. The more robust model seems to be the centisecond Mglob with 74% of good recognition. Concerning the Master/Slave model, the bad results could be explained by a not so good synchronization between the two input signals and once again by the too small database.

4. CONCLUSION

We have proposed an Automatic Speech Recognition System merging visual and acoustic features. As many systems, it is composed of a signals preprocessing and a linguistic decoder. The objective was multiple.

Firstly, we aimed to evaluate the contribution of articulatory features in an ASRS. Secondly, we wanted to test a new method based on a Master/Slave HMM. Then we have compared a centisecond approach with a segmental one both using the « pseudo-diphone » as elementary unit.

Results have shown that articulatory features bring an important increase of the recognition rates whatever the conditions. The segmental approach allows the input information to be greatly reduced. So, in spite of some recognition rates lower than those produced by the centisecond approach, this is an interesting method. Finally, the Master/Slave Model has an architecture enabling to manage asynchronous inputs. But it entails the learning of a great amount of parameters. So, its scores are sometimes worse than those of the global model: it seems to be involved mainly by the use of a too small database for learning correctly a lot of parameters.

5. REFERENCES

1. P.Duchnowski, Meier U., Waibel A. : 'See me, hear me : integrating automatic speech recognition and lip-reading'. S11-6.1 ICSLP94, Yokohama.

2. T.Watanabe, Khoda M. :'Lip-reading of Japanese Vowels Using Neural Networks'. ICSLP90, pp1373-1376, Kobe.

3. P.Jourlin :'Automatic bimodal speech recognition'. ICPhS95, Stocklom.

4. F.Brugnara, De Mori R., Guiliana D., Omologo M. : 'A family of Parallel Hidden Markov Models'. ICASSP92, San-Francisco.

5. R. André-Obrecht, Jacob B., Sénac C. :'Words on Lips : How to Merge Acoustic and Articulatory Information to Automatic Speech Recognition', EUPSICO96, Trieste september 96.

6. C.Abry, Boë L.J. :'Laws for lips'.Speech Communication', pp97-104, 1986.

7. 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.

8. B.Jacob, R.André-Obrecht: 'Sub-dictionary statistical modeling for isolated word recognition'. Proceedings ICSLP94, pp863-866, vol.2, Yokohama.