INTEGRATING VISUAL AND ACOUSTIC INFORMATIONS IN A SPEECH RECOGNITION SYSTEM BASED ON HMM

P. Jourlin, M. El-Bèze and H. Méloni
Laboratoire d'informatique, Avignon, France

ABSTRACT

In an unprotected sonorous environment, an ever so slightly high level of noise can be a hindrance to the correct perception of the acoustic message. Works on multimodal perception show how visual information leads to a compensation of information losses caused by acoustic noises. This explains why an HMM-based system which is able to take information from both visual and acoustic sources can be interesting for speech recognition in noise.

INTRODUCTION

Two main goals appear in speech recognition research : reduce the number of constraints upon working conditions and improve recognition rates of the systems.

The integration of other information sources is actually more and more taken into consideration. These can be a knowledge on different levels : articulatory, auditory, syntactic, morphological, semantic, and so on.

But we can also contemplate taking into account visual information which accompanies, even determines the sound emission. The lips take part in this speech production process.

So, we can feel free to think that labial movements can help speech recognition improvement.

SITUATION OF THE PROBLEM

The main object of this paper is to create an automatic speech recognition system which can draw benefit from lips-originated information.

It is within the AMIBE PROJECT context (Applications Multimodales pour Interfaces et Bornes Evoluées) for which one of the applications could be the conception of advanced bank interfaces.

In this framework, the acoustic signal may be noisy without the video signal being so. This is why labial movements can help to compensate the loss of information due to noise. Some studies have brought to the fore this contribution amongst various persons as well as some problems :

- The labial anticipation and retention which creates a disynchronization between visual and acoustic information, [1],[4].
- The existence of labial doubles which limits labial recognition to some groups of phonemes,[1] [6], [7].
- Coarticulation effects, the incidence of which upon speech signal have been widely studied, [3].

The systems described below are realized from HTK (HMM toolkit) of the C.U.E.D. (Cambridge University Engineering Department Speech Group), some experiments having need a modification of sources.

HIDDEN MARKOV MODELS

Introduction

We chose to resort to these probability models because their use does not require, at first, a thorough knowledge in the application field. The learning stage enable us to detect automatically significant information from the data to which they are submitted.

Each model is represented by a Markov *source* (see Figure 1), which is a probabilistic automaton of finite states, which means that each transition has a probability to be used and a probability to emit symbols.

[1] Also called visemes

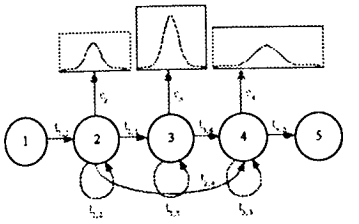

Figure 1 : A Markov source

In the field of speech, these symbols are *N*-dimensional acoustic vectors (ex : LPC of *N* coefficients) but they may also be labial vectors. In the case of a continuous model, the probability emission of these vectors is ruled by a Gaussian mixture.

When some vector components are supposed statistically independents, we can split them into several *streams* (we have one probability distribution for each state and each stream), the calculation of the probability of emission being the product of the probabilities for each stream. The emission and transition probabilities of each model will be estimated from learning data.

Learning stage

We have a set of parametrized and labelled signals at our disposal. The learning stage is composed of 3 stages.

Initialization

Using the *K-means* classification method, we associate to each state of each model, the vectors stemming from the various examples of the concerned word. It needs providing a set of labelled data as an entry of the process.

The start of the learning phase induces a manual intervention. Thanks to the maximum likelihood estimator, we even initialized for each state the means and variances of each component of these vectors, which will enable us to get the associated probability distribution.

This method realizes a rough state-level segmentation as it does not take into account the probabilities of transition and leaves asides the sequential order of the vectors. These various limitations call for the use of a second stage.

Probabilities estimation with temporal constraints

We compute the transition probabilities values by using the Baum-Welch algorithm (also called Forward-Backward). It appears that this stage, based on the initial boundaries of words pains into managing into the questions of contexts related to continuous speech.

This is why it is necessary to proceed a third stage.

Probabilities estimation without temporal constraints

For each sentence of learning data, we create a concatenation of the models corresponding to words pronounced.

Then, we use the Baum-Welch algorithm on the sentence and the associated model, which enables us to take into account the effects related to the context for each element of the concatenation.

This last step enables us to discuss the boundaries of labelling.

Decoding

To recognize the underlying word associated with a given sequence of vectors, we compute for each model the probability that it may have emitted it : the word recognized corresponds then to the model maximizing this last value.

In other words, if O is the sequence of vectors observed, we must compute :

$$\arg\max_i P(w_i \mid O)$$

This value is not computable directly but using Bayes'rule gives :

$$P(w_i \mid O) = \frac{P(O \mid w_i) P(w_i)}{P(O)}$$

w_i being the i-th word of the vocabulary and O the studied sequence of vectors. In the case of continuous speech, we would search for the sequence of models having the highest probability to have emitted this sequence of vectors.

In order to reduce the huge complexity flowing from an optimality exhaustive research, we can apply the *Token passing model* algorithm [10] which is in fact a particular implementation of the Viterbi's algorithm.

THE CORPUS

Sentences are continuously spoken by only one speaker.

The rough data which has been provided to us are the inner-lips height, width and area, synchronized with the acoustic signal [7].

This data set contains two-hundred files, each of which is composed of four letters from A to Z, continuously spoken in french, which will be separated into 70% of files for learning and 30% for test.

THE MODELS USED

We use one model for one word, which enable us to limit the problems of coarticulation and of the setting of boundaries. All the models have the same topology. While it is obvious that results suffer of that choice of topology, (a "A" should not have the same state number as a "W"), it makes the various comparisons and interpretations easier.

The labial model

Realization

The models used in this study are composed of eight states (six emitting states), each one having one transition to the following state and one loop (see Figure 1). Vectors have nine components: height, width, area, their speed and acceleration, divided up into three streams.

Results

We obtain a recognition rate of 42.05% on test data (data which are not learned). These results are surprisingly good, beyond expectation. It must be said on the point that the speaker is particularly cooperative.

The acoustic model

Realization

The chosen parameters are 12 cepstral coefficients spreading on a Mel's scale to which the energy of the signal as well as the derival of each of the former parameters are added.

The sources have the same structure as those used for the labial model, i.e. that all the words have the same topology.

Results

The model, when working on non-noisy data gives us a rate of 87.88% of word recognition.

The bimodal model

Realization

Cesptral coefficients being obtained every 10 ms and labial parameters every 20 ms, an interpolation of the latter is done. This interpolation (linear actually) is a cause of handicap for labial recognition.

The test we have done shows a loss of 10% of recognition between data sampled at 20ms and interpolated data. An alternative should be the use of Lagrange's or Newton's polynomial or splines, interpolations which have not been tested.

Acoustic and labial parameters described above are concatenated, and at last separated into two distinct streams, the topology of both acoustic and labial models is the same.

Results

The most part of information is contained in the acoustic source, so if this one is not noisy, labial information become some kind of noise. This explains the rate of 87.50% of word recognition.

NOISE INFLUENCE UPON RESULTS

We add a crowd noise (vocal frequencies) to the acoustic signal with a sound-noise ratio fixed. For this level of noise we do the learning and the test. Test results are calculated with the following formula :

$$R = \frac{N - I - S - D}{N} \times 100$$

N being a number of units to be recognized, I the number of insertions, S the number of substitutions, D the number of deletions et R the recognition rate.

sound-noise-ratio	labial model	acoustic model	bimodal model
no noise	42.05	87.88	87.50
6 dB	...	73.48	77.65
0 dB	...	53.03	62.88

CONCLUSION

The results we have obtained underline that with a non-noisy signal, labial movements carry some kind of noise, and make fall recognition rates.

However, results obtained in a noisy environment are encouraging and the applications of this type of system are numerous and are not limited to recognition : means, variances and probabilities calculated during the learning stage can be used for synthesis.

Recent studies have underlined that a lips display, in addition with vocal synthesis, considerably improves intelligibility [2].

FUTURE DIRECTIONS

The results we obtained are encouraging but could surely be improved. First, it is essential to increase our data base. It should enable us to have a more accurate learning and more significant tests.

We have to handle disynchronization between visual and acoustic information. It could lead to a modification of decoding algorithms as well as a modification of model structure.

At last, it is necessary to study further the management of weight tied to the various information sources according to various measures : sound environment, phonemes, acoustic or visual features, etc.

REFERENCES

1. C. Abry et M.T. Lallouache (1994) *Pour un modèle d'anticipation dépendant du locuteur - Données sur l'arrondissement en français* - Bulletin de la communication parlée

2. A. Adjoudani, T. Guiard-Marigny, B. Le Goff et C. Benoît (1994) *Un modèle 3D de lèvres parlantes* - 20èmes journées d'étude sur la parole

3. C. Benoît, T. Mohamadi, S.D. Kandel (1994) *Effects of phonetic context on audio-visual intelligibility of french* - Journal of Speech and Hearing Research

4. M-A. Cathiard et M.T. Lallouache (1992) *L'apport de la cinématique dans la perception visuelle de l'anticipation et de la rétention labiale* - 19èmes J.E.P. - Bruxelles

5.M. Gentil et L-J. Boë (1979) *Les lèvres et la parole : Données anatomiques et aspects physiologiques* - Travaux de l'institut de phonétique de Grenoble

6. M. Gentil (1981) *Etude de la perception de la parole : lecture labiale et sosies labiaux* - Etude pour le centre scientifique IBM de Paris

7. M.T. Lallouache (1991) *Un poste "visage-parole" couleur* - Thèse de doctorat - ICP Grenoble

8. E.D. Petajan (1884) *Automatic lipreading to enhance speech recognition* - Proceedings of the Global Communications Conference, IEEE Communication Society, Atlanta, Georgia, 265-272.

9.J. Robert-Rives (1995) *Modèles d'intégration audio-visuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique de voyelles* - Thèse de doctorat - ICP Grenoble

S.J. Young, N.H. Russel, J.H.S. Thornton (1989) *Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems* - Technical report - CUED/F-INFENG/TR.38