COMPARATIVE STUDY BETWEEN TWO RECOGNITION METHODS CHMM AND SVM FOR AUTOMATIC RECOGNITION OF AUDIOVISUAL SPEECH

Nadia Bakir
University of Science and Technology Houari Boumediene (USTHB)
Speech and Signal Processing Laboratory
Faculty of Electronics and Computer Science
bak_nad03@yahoo.fr

***Abstract*—** In this paper, we present a comparative study between two recognition methods for the same system of automatic recognition of audiovisual speech (ARAVS) combining acoustic data and visual data. These AAVSR systems use as methods of recognition the continuous hidden Markov models CHMM (Continuous Hidden Markov Model) and support vector machines (SVM) or wide margin separator. And as a method of fusion, the Separated Identification (SI) based on an MLP (PMC) NN. The visual information used in conjunction with the acoustic data is based on the shape and movements of the lips during speech. The experiments carried out for the recognition of the Arabic numerals indicate that the association of the acoustic modality and the visual modality increases the performance of the system of Automatic Speech Recognition (ASR) in real environment (highly noisy), an increase in the Recognition Rate (RR) of 17% for CHMM and 15% for SVM was noted.

Keywords— *ASR, Audiovisual fusion, Continuous HMM, SVM, Neural Networks.*

I. Introduction

The use of additional information in conjunction with that extracted from the acoustic signal is a new method used to improve the performance and robustness of automatic speech recognition systems. The use of data based on the shape and motion of the speaker's lips therefore seems to be a promising way for speech recognition.

For this purpose and through this work, we are interested in the extraction of visual information, and methods of integrating acoustic and visual scores using two different recognition or classification methods (CHMM or SVM). (CHMM or SVM).

II. Speech recognition system by audio-visual fusion

The structure of the Audio-Visual Fusion Recognition System (SRFAV) implemented is given in Figure 1 below. This system comprises three modules which are: the acoustic recognition module, the visual recognition module and the fusion module.

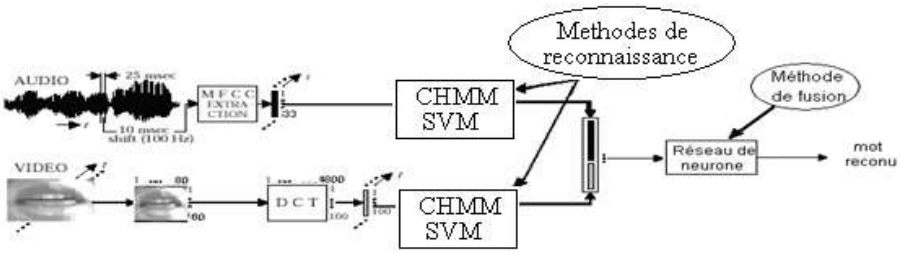

Figure 1: Structure of the Audiovisual Fusion Recognition System.

The acoustic recognition module uses two recognition methods, the first is stochastic and based on continuous hidden Markov models (Rabiner, 1989) (Boite, 1999), and the second method is a statistical classifier based on SVM (Platt, 1998) (Callut, 2002).

Its generic process is based on three phases that are:

the parameterization of the acoustic signal, using in our case, the MFCC coefficients (Mel Frequencies Cepstral Coefficients), learning models or classes, and the decoding phase based on the Viterbi algorithm (Rabiner, 1989) (Boite, 1999) for CHMMs and on the algorithm of SMO (Sequential Minimal Optimization) for SVM (Platt, 1998).

The visual recognition module uses the same methods; it differs only in its parameterization phase based on the DCT (Discrete Cosine Transform) (Bovik, 2000). The fusion module is based on the Separate Identification (SI) fusion method using arrays of multilayer perceptron artificial neurons.

II.1. Data processing

II.1.1. Audiovisual separation

Once the recording of the video sequences of the speaker is made using the Windows Movie Maker software with the extension " .wmv ", the first operation is to convert the video sequences of the extension '.wmv' to the extension '.avi', using the BPS (Video Converter & Decompiler) software. Then, we move on to the separation of the two audio and video streams.

The audio stream is extracted as a signal using the Gold Wave software with the " .wav " extension, and from the video stream using the BPS software, we extract still images from the sequence. Then we go on to build the two audio and video databases.

II.1.2. Acoustic data

The processing of acoustic data is based on the calculation of the spectral envelope through the cepstral coefficients in the Mel scale, namely the MFCC coefficients (Mel Frequencies Cepstral Coefficients) (Hermansky, 1990). The acoustic signal is thus first filtered through a filter

with the following transfer function:

$$H(Z) = 1 - 0.95 \times Z^{-1} \quad (1)$$

It is then fragmented into frames and weighted by a Hamming window of 25.6ms with a displacement of 10ms. To the MFCC coefficients is added the dynamic information carried by the speed (ΔMFCC) and the acceleration (ΔΔMFCC) (Wilpon, 1993). Each frame is thus represented by a vector xt of the following form:

$$x_t = \{MFCC(m), \Delta MFCC(m), \Delta\Delta MFCC(m)\} \quad (2)$$

II.1.3. Visual data

To characterize the video signals we use the DCT (Discrete Cosine Transform), whose coefficients are classified in the increasing order of the low frequencies. This technique is widely used in JPEG or MPEG image or video encoders (Bovik, 2000). The DCT is similar to the Fourier transform because it transposes the time domain into the frequency domain. On the other hand, unlike the Fourier transform, it only includes the real coefficients.

The two-dimensional cosine transform (DCT) and the inverse transform (IDCT) are defined by the following relationships 3 and 4:

$$H(u,v) = \frac{2}{\sqrt{MN}} C(u)C(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} h(x,y) \cos\left[\frac{(2x+1)u\pi}{2M}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (3)$$

$$h(x,y) = \frac{2}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} C(u)C(v)H(u,v) \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (4)$$

with :

$$C(u) = \begin{cases} \frac{1}{\sqrt{2}} \; pour \quad : u = 0 \\ 1 \; pour \quad : u > 0 \end{cases} \quad (5)$$

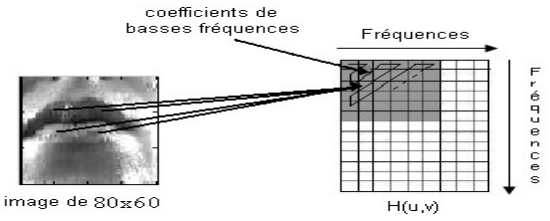

Figure 2: Graphical representation of the DCT.

The input vectors are formed of the low frequency coefficients that are in the upper left corner of the resulting matrix, as shown in Figure 2. In this figure, we retain only the first 100 high-amplitude coefficients of an image of dimension 80x60 (4800 coefficients), so the visual vector is in this case is composed of 100 elements.

The number of high-amplitude coefficients retained after the transformation by the DCT is chosen so as to keep at least 80% of the total energy in the high-amplitude coefficients which will be sufficient to reconstruct the main characteristics of the image. The total energy of the image is calculated according to Parseval's theorem from the coefficients of the DCT by the following relation 6:

$$E = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} |H(u,v)|^2 \quad (6)$$

Figure 3: Reconstitution of an image from 100 high amplitude coefficients of dimension 80x60.

The main idea of the algorithm to encode the image by the DCT is not to use all the coefficients (4800 coefficients), in order to limit the memory size and the calculations necessary for the training and the recognition by the two modules or methods of recognition CHMM and SVM. In our work, we have kept the first hundred (100) coefficients to represent the image.

II.2. Recognition methods

In the ASR domain, the signals are coded as temporal vibrations of short duration spectrum; for this, we present in this work the two methods of recognition CHMM and SVM that can be used for handling problems in which the information is uncertain and incomplete.

II.2.1. Continuous Hidden Markov Models

The recognition module based on Continuous Hidden Markov Models (CHMM) is used to recognize the acoustic modality as well as the visual modality. Continuous models use continuous probability density functions to evaluate probabilities of observation directly in the primitive space. Each state models its observations independently of the other states of the model, by a weighted sum of elementary functions. The probability density function given by equation 7, associated with states, is a weighted sum of Gaussian Mixture Model (GMM), it is given by equation 8:

$$b_j(x_t) = \sum_{m=1}^{M} C_{im} \times N(x_t; \mu_{im}; \Sigma_{im}) \quad (7)$$

or:

$$N(x_t; \mu_{im}; \Sigma_{im}) = \frac{1}{(2\pi)^{\sigma/2} |\Sigma_{im}|^{\frac{1}{2}}} exp\left(-(x_t - \mu_{im})^T \sum_{k,j}^{-1} (x_t - \mu_{im})\right) \quad (8)$$

With μim and Σim are respectively the mean vector and the covariance matrix of the Gaussian mth of the state i, cim the weighting coefficient assigned to it and M the number of Gaussian

mixtures.

The model chosen in Figure 4 is a left-right model with three emitter states.

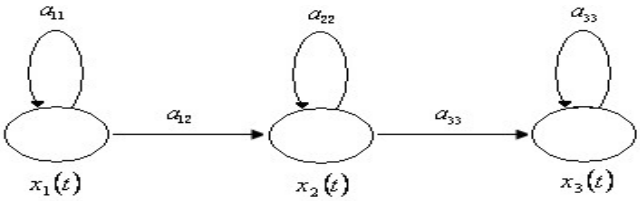

Figure 4: Left-right model with three emitting states.

II.2.2. Support vector machine

The recognition method based on Support Vector Machines (SVM) is used to recognize both acoustic and visual modalities. The task of this recognizer is expressed by a function called a decision function, linking the examples to classify x (called input space) to their class y (called output space). Hence, y often corresponds to {-1, + 1}.

Is: $\{f_\alpha, \alpha \in \Lambda\}, f_\alpha : R^d \longrightarrow \{\pm 1\}$.

The set of functions, such that: $(x_1, y_1), \ldots, (x_l, y_l) \in R^d \times \{\pm 1\}$, are independent learning examples, generated randomly according to an unknown probability distribution P (x, y).

In our case, the machine is supposed deterministic: for a given x and a given α, we always get the same output f {x, α}. A particular choice of α generates what we will call "trained machine".

The objective of the SVMs is to find a hyper-plane that separates the learning data so that all the examples of the same class are on the same side of the hyper-plane (Hsuand , 2002) (Callut, 2002).

Originally, SVMs were designed for the separation of two classes. However, two approaches allow extending this algorithm to the case of several classes (Hsuand , 2002), the One Against All (OAA) approach that we will use in this work, and the One Against One (OAO) approach.

The synoptic diagram of the OAA system is given in figure 5:

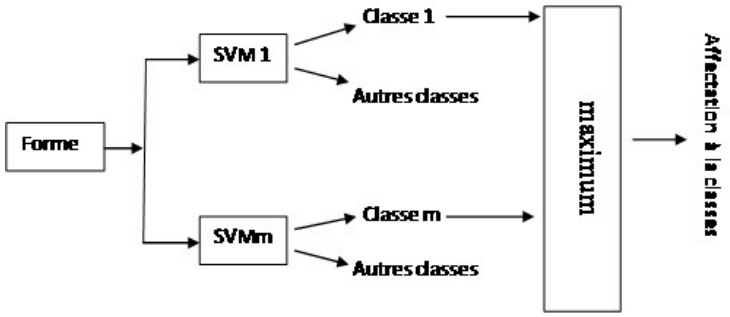

Figure 5: Synoptic diagram of the One Against All 'OAA' system.

II.3. Audiovisual fusion

The integration of auditory and visual information can be done in different ways (Adjoudani, 1993) (Rogozan, 1999) (Potamianos, 2004) (Potamianos, 2003). Fusion models are classified into three main categories: the direct fusion model, the separate fusion model and the hybrid fusion model (Wilpon, 1993). In this work, we used the separate fusion model or more precisely the fusion of scores from each recognizer (acoustic and visual recognizer). This fusion is achieved through an MLP (Multi Layer Perceptron) neural network, whose architecture is shown in figure 6 below. It is a network with three layers consisting of an input layer containing 20 cells; a hidden layer of 100 cells and an output layer of 10 cells.

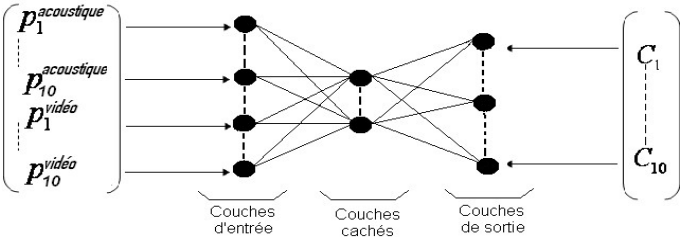

Figure 6 :Fusion network architecture.

III. Experiences and results

III.1. Database

In this first work, a database in mono-speaker mode was built.

This database was recorded in a real conditions. It includes pronunciations of isolated Arabic numerals taken at a sampling frequency of 16 KHz. It consists of 25 repetitions for each Arabic numeral from zero to nine (siffer, wahed, ithnani, thalatha, arbaa, khamssa, sitta, sabaa, thamania, tissaa).

III.2. Audiovisual results

For all the training and recognition tests performed, the algorithm parameters for both methods were chosen to obtain the best results.

The obtained results are summarized in three tables. Table 1 shows for each digit the acoustic, visual and audiovisual Recognition Rates ‘‘RR’’, for both recognition methods, for the Base Test. Table 2 summarizes the performances expressed for acoustic, visual and audiovisual systems in terms of RR Global (RRG) for the two training and test Bases. Table 3 summarizes the overall comparative results obtained between the two recognition methods CHMM and SVM.

Table 1: Recognition Rate RR in (%) for the audio, video and audio visualcases. two methods of recognition are employed.

	Recognizer CHMM			**Recognizer SVM**		
Digits	Results A (%)	Results V (%)	Results A (%)	Results A (%)	Results V (%)	Results AV (%)
Siffer	86,66	73,33	90	66,66	80	93
Wahed	66,66	60	85	73,33	80	85
Ithnani	53,33	46,66	87	46,66	60	75
Thalatha	73,33	60	89	53,33	66,66	70
Arbaa	53,33	46,66	83	46,66	53,33	72
Khamsa	60	53,33	89	66,66	73,33	82
Sitta	66,66	60	88	60	73,33	81
Sabaa	80	53,33	85	66,66	66,66	76
Thamania	80	66,66	89	86,66	73,33	75
Tissaa	73,33	73,33	85	80	80	90
RRG	69,33	59,33	87	64,662	70,664	79,9

Discussion 1

Table 1 shows the evaluation of our system for each digit.

The acoustic recognition rate RR is higher than the visual RR for both recognition methods, which means that the acoustic system is better than the visual one. On the other hand, the best system is the audiovisual system as seen in the results of the audiovisual RR.

Table 2: Comparison between the two methods of recognition chmm and svm.

	Recognizer CHMM			**Recognizer SVM**		
	RRG	RRG	RRG	RRG	RRG	RRG
Results	A(%)	V(%)	AV(%)	A(%)	V(%)	AV(%)
BA	94	87	99	90	96	98
BT	69,33	59,33	87	64,662	70,664	79,9

Discussion 2

Table 2 summarizes the overall comparative results of the Automatic Recognition experiments by the Audio-Visual Speech ARAVS system using the two recognition approaches CHMM and SVM. Note that the RRG of the CHMM is greater than the RRG of the SVM for the acoustic modality. On the other hand, the RRG of the CHMM is lower than the RRG of the SVM for the visual modality. That is, the CHMM appears to be the best performing approach for acoustic ASR, and the SVM is the most powerful approach for visual ASR.

We also note that the audiovisual RRG is higher than the acoustic and visual RRG, which shows that the audiovisual association, based on the Separate Identification model SI, can improve the ASR performances.

IV. Conclusion

The main objective of this work is the implementation of an automatic recognition system of audio-visual speech by two methods of recognition. Thus, we tackled the fusion of the acoustic and visual information for the task of ASR. Thus, an audiovisual integration model with separate identification has been developed. The system implemented was tested on an audiovisual corpus, consisting of sequences of mono-speaker Arabic numerals, in a very noisy environment.

Tests have shown that the integration of the visual modality based on the separate fusion model improves the performances of the recognition system in noisy environment : RR = 69.33% for the audio system alone by CHMM), RR = 59.33%, for the visual system by CHMM, and RR = 87%, for the audiovisual system by CHMM. The SVM appears to be a good classification method applied to RAPAV too.

In perspectives, we are interested in the use of larger databases uttered by multiple speakers to test the real reliability of our system.

We also plan to test our system in the case of corrupted speech signals by specific noises and at different SNR levels (Signal on Noise Ratio) using other methods of recognition as the Hybrid HMM / SVM or fusion methods such as direct fusion and hybrid fusion.

References

Adjoudani Ali., 1993. “Élaboration d'un modèle de lèvres 3D pour animation en temps reel”, Mémoire de D.E.A, Signal Image Parole, Institut National Polytechnique de Grenoble, France, 1993.

Boite R., 1999, H. Bourlard, T. Dutoit, J. Hancq et H. Leich, “Traitement de la parole”, Collection Electricité, presses polytechniques et universitaires remandes,EPLF-Centre Midi, CH-1015 Lausanne. 1999.

Bovik A., 2000. ‘’Handbouk of Image and Video Processing’’, AcademicPress,p891. 2000

Callut J., 2002. “Implementation efficace des Support Vector Machine pour la classification”. Mémoire présenté en vue de l’obtention de grade du maitre en informatique. Université Libre de Bruxelles, pp. 108, Belgique, 2002.

Hermansky H., 1990. “Perceptual linear predictive analysis of speech”, Journal of the Acoustical Society of America, Vol. 87, N° 4, pp. 1738-1752, 1990.

Hsuand C.W., 2002, C., J. Lin, “Comparison of methods for multiclass support vector machines”. IEEE, Transactions on Neural Networks, Vol. 13, pp.415-425. 2002.

Platt J. C., 1998. “Fast training of support vector machines using Sequential Minimal Optimisation” ; SMO-Book, chap. 12, pp.41-65, Scholkopf, C. J. C. Burges, A. J. Smola, editors, Advance in Kernel Methods – Support Vector

Learning, MA,. 1998.

Potamianos G., 2003, C. Neti, S. Deligne, “Joint Audio-Visual Speech Processing for Recognition and Enhancement”. IEEE, proceeding of the Auditory-Visual Speech Processing Tutorial and Research Workshop (AVSP), pp. 95-104, St, Jorion France, September 2003.

Potamianos G., 2004, C. Neti, J. Luttin, I. Matthews, “Audio-visual automatic speech recognition: an overview”. In issues in audio-visual speech processing (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004.

Rabiner L. R., 1989. “A tutorial on Hidden Markov Models and selected applications in speech recognition”, Proceeding of the IEEE, vol. 77(2), 1989.

Robert-Ribes J., 1995. “Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique des voyelles”, Thèse de doctorat, Signal Image Parole, Institut National Polytechnique de Grenoble, France, 1995.

Rogozan A., 1999. “Etude de la fusion des données hétérogènes pour la reconnaissance automatique de la parole audiovisuelle”, Thèse PHD, Ecole doctorale en électronique de l’université d’Orsay, Paris, 1999.

Wilpon J.C., 1993, C. H. Lee, L.R. Rabiner, “Connected digit recognition based on improved acoustic resolution” Computer Speech and Language, 7, pp 15-26, 1993.