Audio-Visual Speech Recognition One Pass Learning with Spiking Neurons

Renaud Séguier and David Mercier

Supélec, Team ETSN
Avenue de la Boulaie, BP28
35511 Cesson Sévigné, France
{Renaud.Seguier, David.Mercier}@supelec.fr
http://www.supelec-rennes.fr/ren/rd/etsn/

Abstract. We present a new application in the field of impulse neurons: audio-visual speech recognition. The features extracted from the audio (cepstral coefficients) and the video (height, width of the mouth, percentage of black and white pixels in the mouth) are sufficiently simple to consider a real time integration of the complete system. A generic preprocessing makes it possible to convert these features into an impulse sequence treated by the neural network which carries out the classification. The training is done in one pass: the user pronounces once all the words of the dictionary. The tests on the European M2VTS Data Base shows the interest of such a system in audio-visual speech recognition. In the presence of noise in particular, the audio-visual recognition is much better than the recognition based on the audio modality only.

1 Audio-Visual Speech Recognition

Speech recognition in noisy environments is useful in many applications, for example in car computer vocal interface or automatic ticket sale in stations and airports. The significant contribution of information contained in the movement of the lips makes it possible to improve the audio recognition rates. Audio-visual speech recognition systems use mainly HMM [3]. We propose in this article such a system exploiting impulse neurons (STAN Spatio-Temporal Artificial Neurons [15] [8]) and allowing a light training since it is only carried out on one pass. The Audio-visual features are sufficiently simple and robust to consider a real time implementation on low quality audio and video signals such as those provided by usual webcams.

2 Proposed System

The system is illustrated in Figure 1. After a specific preprocessing performed separately on the audio and video signals, a generic preprocessing makes it possible to produce impulse sequences taken into account by the STAN's which operates the classification.

J.R. Dorronsoro (Ed.): ICANN 2002, LNCS 2415, pp. 1207–1212, 2002.

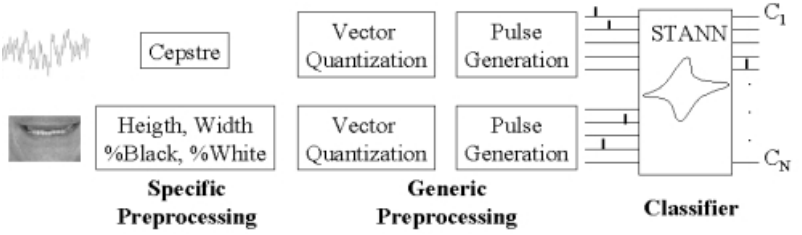

Fig. 1. Audio-Visual Speech Recognition System.

2.1 Specific Preprocessing

Audio. Cepstral coefficients are often used to characterize the sound in this type of application [6] [4]. On a 40ms sliding window, we calculate as [7] the first 12 coefficients of the cepstrum, the logarithm of the signal energy in the window and the temporal derivate of those thirteen parameters. Each one of them is normalised taking into account the values which are observed all along the sequence (from 0 to 9).

Video. Certain teams carry out a PCA (Principal Component Analysis) [1] or a DCT (Discrete Cosine Transform) [10] but most of the time dynamic contours [2] or deformable models [14] are used to characterize the shape of the mouth. Our objective is to make a real time system, thus we use features which are much simpler to extract. The height and the width of the mouth are evaluated with the method presented in [11]. For the height evaluation of the mouth, instead of working on grey levels, we use V values (from YUV color coordinate system): in these coordinates, the teeth and the dark interior of the mouth are confused, which enables us to locate the upper and lower lips more precisely.

We define in addition a sub-mouth area (see Fig. 3) centered on the mouth, in which we calculate an eight-bit grey level histogram of the pixels. The ratio between the number of pixels whose values are lower than 50 and the total pixel number gives us the percentage of dark pixels, that between the number of pixels whose values are superior to 150 and the total pixel number gives us the percentage of light pixels. We add to these four parameters (height, width, dark percentage, light percentage), the temporal derivate of the height and the width. As for the audio, we normalize each parameter over the sequence from 0 to 9.

2.2 Generic Preprocessing

We use a generic preprocessing [13] in order to convert the temporal series of features into an impulse sequence. This stage consists in applying a vector quantization (K-means [5]) separately on the audio and video features in order to extract vectors codes. At each instant, the impulse generation module compares

the Euclidian distance between signal (cepstral coefficients or mouth features) and these code vectors. Each output of the impulse generation module characterizes a vector code. An impulse is then generated on the output associated with the code vector which is closest to the input signal.

2.3 Classifier

The STAN (Spatio-Temporal Artificial Neuron) works in the complex domain. An impulse sequence is converted into a vector X with complex values in the following way (Fig. 2).

The impulse of amplitude η_1 emitted at time t_1 on component j is coded at current time t by the complex number:

$$x_j(t) = \eta_1 \, exp[-\mu_S \tau_1] \, exp[-i \arctan \mu_T \tau_1] \tag{1}$$

with $i = \sqrt{-1}$, $\tau_1 = t - t_1$ and $\mu_S = \mu_T = \frac{1}{TW}$

TW depends on the application and represents the size of the temporal window inside which impulse sequence must be identified. When a new impulse η_2 is emitted at time t_2 on the same component, it is accumulated in the component j of the vector X:

$$\begin{aligned} x_j(t_2) &= \eta_1 \, exp[-\mu_S (t_2 - t_1)] \, exp[-i \arctan \mu_T (t_2 - t_1)] + \eta_2 \\ &= \rho e^{i\phi} \end{aligned} \tag{2}$$

and later:

$$x_j(t) = \rho \, exp[-\mu_S (t - t_2)] \, exp[-i \arctan (tan\phi + \mu_T (t - t_2))] \tag{3}$$

Each component of the vector X is thus reactualized as soon as an impulse is presented to the input. The comparison between X and the weight W of the neuron itself characterized by a complex vector is done here by the means of a Hermitian distance D:

$$D(X, W) = \sqrt{\sum_{j=1}^{N} (x_j - w_j)\overline{(x_j - w_j)}} \tag{4}$$

knowing that $\overline{x}$ is the complex conjugate of x.

Learning Phase. There are as many neurons in the output layer as words to be recognized. Each neuron is characterized by a weight vector. The training is done in one step only. It consists in presenting as input the audio-video sequence corresponding to the word which we wish to recognize. An impulse sequence is then generated and converted into a complex vector X (see Fig. 2) which characterizes the presented word. We carry out this procedure on the whole dictionary in order to evaluate each vector having to characterize each word of the dictionary. These vectors define each weight vector W of the STAN's used during the classification.

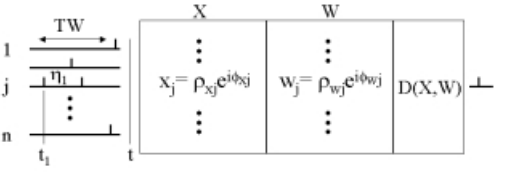

Fig. 2. The Spatio-Temporal Artificial Neuron (STAN)

Testing Phase. When an unknown audio-video sequence is presented at the input, it is translated in the form of an impulse sequence, converted into a complex vector and compared to each weight by the means of the Hermitian distance. The neuron producing the minimal distance then emits an impulse at the output: it signals the recognized word.

3 Results and Conclusion

3.1 Tests

M2VTS. Within the framework of separated word recognition, we tested our system on the first ten persons of the European Data Base M2VTS [9] (Multi Modal Checking for Teleservices and Security applications). This base is dedicated to Audio-visual recognition and identification. Each person pronounces four times (at one week interval) the digits from 0 to 9. We chose this base because it characterizes well the conditions of use in which the real time implementation of our system will have to function. The images were acquired at 25Hz with a weak resolution (288x360pixels in 4:2:2), the sound was sampled at 48kHz on 16 bits. Some people smile sometimes during acquisition, which considerably harms the performance of lipsreading.

System Parameters. We use a face detector [12] in order to locate and follow the face during the sequence. An evaluation of the motion inside the face enables us to locate the mouth rather precisely as shown in Figure 3. In the interior of the mouth, we delimit a zone of 11 pixels height in which the percentages of dark and light pixels are evaluated. With regard to the parameter setting of the STAN's, we used the same value of TW (ten units which correspond to an 400ms observation window). Thirty vectors codes were extracted from the sound, eighty from the video.

Results. Let us recall that we tested our system in the framework of separated word monospeaker recognition. For each person, we have four sequences during which the digits from 0 to 9 are pronounced. Four evaluations were thus carried out according to the number of the sequence used for the training, tests being performed on the three other sequences.

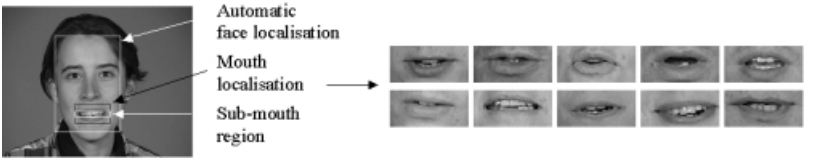

Fig. 3. Automatic Face and mouth localisation.

As one can notice on Figure 4 the performances of the combined audio-visual system (86%) are 10% higher than that of the audio system alone (76%) although the recognition capacities of the video system are definitely lower (61%). But it is in the presence of noise (white noise added to the sound signal) that the audio-visual recognition system takes all its interest. For a signal to noise ratio of 10dB for example, the performances of the combined audio-visual system (68%) are better by almost 20% compared with that of the audio system alone (49%).

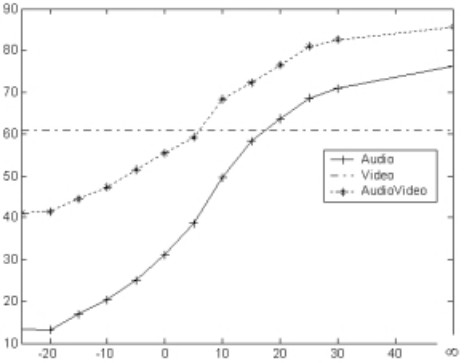

Fig. 4. Pourcentage of correct classification versus Signal to Noise ratio

3.2 Discussion

This work shows that a one pass learning system can extract sufficient information from a learning base to carry out a rather relevant monospeaker speech recognition.

The following stage will consist in conceiving a system recognition of same type as that proposed by [6] who makes a training on the whole (but one) of the people present in M2VTS Data Base and tests on the unused person.

Our final objective is to conceive an "unknown speaker" recognition system which could specialize itself on a particular person, without having to define a

specific training phase. At the moment, we analyse the STAN's output to give us a confidence estimation of the recognition. When this output is strong in the case of a word pronounced by an unknown person, we would like to take into account the input signal to modify the STAN's weights and thus realise an automatic phase of specialization.

References

1. P. de Cuetos, N. Chalapathy, and W. Andrew. Audio-visual intent-to-speak detection for human-computer interaction. In *ICASSP*, 2000.
2. P. Delmas, P.Y. Coulon, and V. Fristot. Automatic snakes for robust lip boudaries extraction. In *ICASSP*, 1999.
3. S. Dupont and J. Luettin. Audio-visual Speech modeling for continuous speech recognition. *IEEE Transactions on multimedia*, 2000.
4. S. Durand and F. Alexandre. Learning Speech as acoustic sequences with the unsupervised model, tom. In *NEURAP, 8th International Conference on Neural Networks and their Applications*, 1995.
5. A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer Acad. Pub., 1991.
6. J. Luettin. Visual Speech and speaker recognition. In *PhD Dissertation, Univ. of Sheffield*, 1997.
7. D. Mercier and R. Séguier. Spiking neurons (stanns) in speech recognition. In *3rd WSEAS International Conference on Neural Network and Applications*, Feb 2002.
8. N. Mozayyani, A. R. Baig, and G. Vaucher. A fully neural solution for on-line handwritten Character recognition. In *IJCNN*, 1998.
9. S. Pigeon. M2vts. In *www.tele.zacl.ac.be/PROJECTS/ M2VTS/m2fdb.html*, 1996.
10. Gerasimos Potamianos and Chalapathy Neti. Automatic speechreading of impaired Speech. In *Audio-Visual Speech Processing*, September 2001.
11. R. Séguier, N. Cladel, C. Foucher, and D. Mercier. Lipreading with spiking neurons: One pass learning. In *International Conference in Central Europe on Computer Graphits, Visualization and Computer Vision*, Feb 2002.
12. R. Séguier, A. LeGlaunec, and B. Loriferne. Human faces detection and tracking in video sequence. In *Proc. 7th Portuguese Conf. on Pattern Recognition*, 1995.
13. R. Séguier and David Mercier. A generic pretreatment for spiking-neuron. Application on lipreading with stann (spatio-temporal artificial neural networks). In *International Conference on Artificial Neural Networks and Genetic Algorithms*, 2001.
14. Y. Tian, T. Panade, and J. F. Cohn. Recognizing action units for facial expression analysis. IEEE Trans. ora Patterra Analysis and Machine Iatelligence, 2001.
15. G. Vaucher. An algebraic interpretation of psp composition. In *BioSystems*, Vol 48, 1998.