BIMODAL RECOGNITION OF ITALIAN PLOSIVES

P. Cosi, M. Dugatto, F. Ferrero, E. Magno Caldognetto and K. Vagges
Centro di Studio per le Ricerche di Fonetica (CNR)

ABSTRACT

A bimodal automatic speech recognition system, in which the speech signal is synchronously analyzed by an audio channel producing spectral-like parameters every 2 ms and by a visual channel computing lip and jaw kinematic parameters, is described and some results are given for various speaker independent phonetic recognition experiments regarding the Italian plosive class in different noisy conditions.

INTRODUCTION

Audio-visual automatic speech recognition (ASR) systems can be conceived with the aim of improving speech recognition performance, mostly in noisy conditions [1]. Various studies of human speech perception have demonstrated that visual information plays an important role in the process of speech understanding [2], and, in particular, "lip-reading" seems to be one of the most important secondary information sources [3]. Moreover, even if the auditory modality definitely represents the most important flow of information for speech perception, the visual channel allows subjects to better understand speech when background noise strongly corrupts the audio channel [4]. Mohamadi and Benoît [5] reported that vision is almost unnecessary in rather clean acoustic conditions (S/N > 0 dB), while it becomes essential when the noise highly degrades acoustic conditions (S/N ≤ 0 dB).

METHOD

The system being described takes advantage of jaw and lip reading capability, making use of a new system for automatic jaw and lips movement 3D analysis called ELITE [6], in conjunction with an auditory model of speech processing [7] which have shown great robustness in noisy condition [8].

The speech signal, acquired in synchrony with the articulatory data, is prefiltered and sampled at 16 KHz, and a joint synchrony/mean-rate auditory model of speech processing [7] is applied producing 80 spectral-like parameters at 500 Hz frame rate. In the experiments being described, spectral-like parameters and frame rate have been reduced to 40 and 250Hz respectively in order to speeding up the system training time. Input stimuli are segmented by SLAM, a recently developed semi-automatic segmentation and labeling tool [9] working on auditory model parameters.

The visual part of the system has adopted ELITE which is a fully automatic movement analyzer for 3D kinematic data acquisition. This system ensures a high accuracy and minimum discomfort to the subject. In fact, only small, non obtrusive, passive markers of 2mm of diameter, realized by reflective paper, are attached onto the speaking subject's face. The subjects are placed in the field of view of two CCD TV cameras at 1.5 meters from them. These cameras light up the markers by an infrared stroboscope, not visible in order to avoid any disturbance to the subject. ELITE is characterized by a two level architecture. The first level includes an interface to the environment and a fast processor for shape recognition (FPSR). The outputs of the TV cameras are sent at a frame rate of 100 Hz to the FPSR which provides for markers recognition based on a cross-correlation algorithm implemented in real-time by a pipe-lined parallel hardware. This algorithm allows the use of the system also in adverse lighting conditions, being able to discriminate between markers and reflexes of different shapes although brighter. Furthermore, since for each marker several pixels are recognized, the cross-correlation algorithm allows the computation of the weighted center of mass increasing the accuracy of the system up to 0.1mm on 28cm of field of view. The coordinates of the recognized markers are sent to the second level which is constituted by a general purpose personal computer. This level provides for 3D coordinate reconstruction, starting from the 2D perspective projections, by means of a stereophotogrammetric

procedure which allows a free positioning of the TV cameras. The 3D data coordinates are then used to evaluate the parameters described hereinafter.

Finally both audio and visual parameters, in a single or joint fashion, are used to train, by means of the Back Propagation for Sequences (BPS) [10] algorithm, an artificial Recurrent Neural Network (RNN) to recognize the input stimuli.

A block diagram of the overall system is described in Figure 1 where both the audio and the visual channel are shown together with the RNN utilized in the recognition phase.

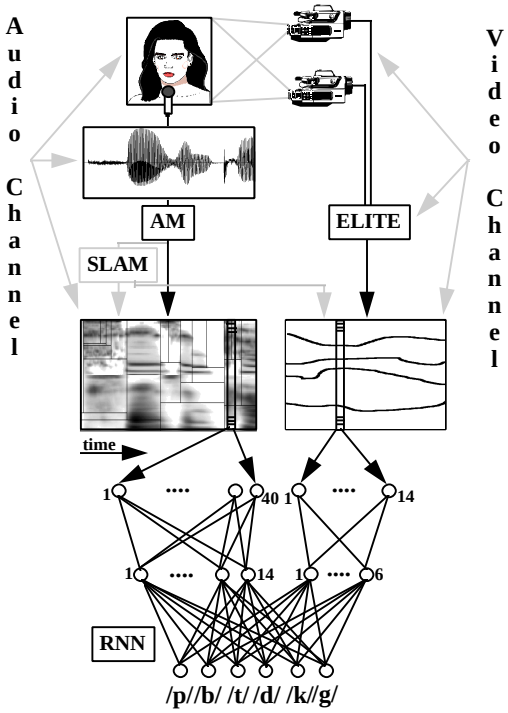

Figure 1. Structure of the bimodal recognition system.

EXPERIMENT

The input data consist of disyllabic symmetric /'VCV/ nonsense words, where C=/p,t,k,b,d,g/ and V=/a,i,u/, uttered by 10 male speakers. All the subjects were northern Italian university students, aged between 19 and 22, and were paid volunteers. They repeated five times, in random order, each of the selected non-sense words. The speaker comfortably sits on a chair, with a microphone in front of him, and utters the experimental paradigm words, under request of the operator. Three reference points and five target points on the face of the subject have been considered. As illustrated in Figure 2, these points are the nose (n.1), the middle edge of the upper lip (n.2), the middle edge of the lower lip (n.5), the corners of the mouth (n.3 and n.4), the chin (n.6), and the lobe of the ears (n.7 and n. 8).

In this study, the movements of the markers placed on the central points of the vermilion border of the upper lip (marker 2) and lower lip (marker 5), together with the movements of the marker placed on the edges of the mouth (markers 3, 4), were analyzed, while the markers placed on the tip of the nose (marker 1), and on the lobe of the ears (markers 7, 8), served only as reference points. In fact, in order to eliminate the effects of the head movement, the opening and closing gestures of the upper and lower lip movements were calculated as the distance of the markers 2 and 5 placed on the lips, from the plane depicted in Figure 2 and defined by the line passing from the markers 7 and 8, placed on the ear lobes, and marker 1, placed on the tip of the nose. Similar distances with a plane perpendicular to the above one serve as a measure of upper and lower lip protrusion. A total of 14 values, defined as the difference between various markers or between markers and reference planes, plus the correspondent instantaneous velocity, obtained by numerical differentiation, constitute the articulatory vector which has been used together with the acoustic vector in order to represent the target stimuli. The articulatory parameters were, besides the upper and lower lip opening and closing movements, and the upper and lower lip protrusion, the lip opening height calculated as the distance between markers 2 and 5, the lip opening width, calculated as the distance between markers 3 and 4, and the jaw opening measured between the markers placed on the jaw and on the tip of the nose.

As an example of the articulatory parameters, Figure 3 shows the vertical displacement and the instantaneous velocity of the marker placed on the lower lip (n. 5) associated with the sequence /'apa/.

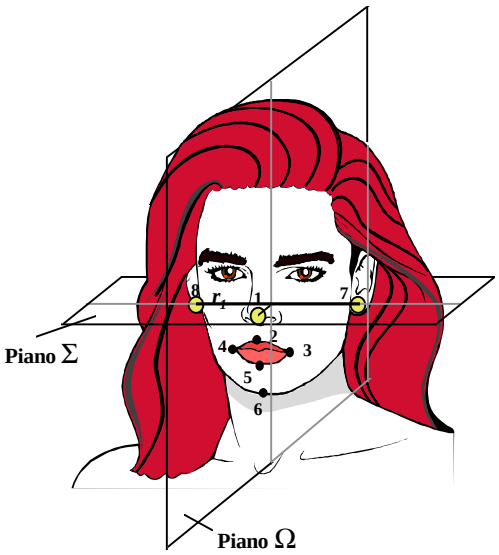

Figure 2. Position of the reflecting markers and of the reference plane. Identification numbers are indicated next to their corresponding markers. Marker dimension in the figure does not correspond to the real dimension (2mm) but is exaggerated for visualization purpose.

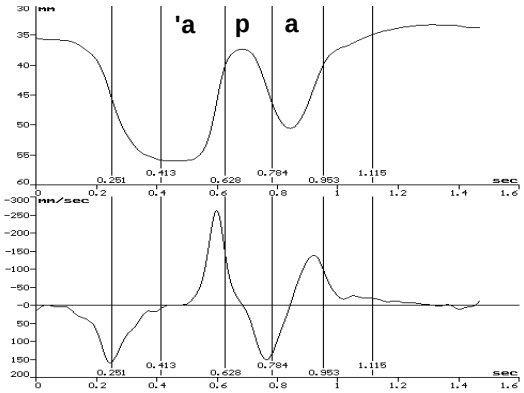

Figure 3. Time evolution of displacement and velocity of the marker placed on the lower lip (n.5), associated with the sequence /'apa/.

In a previous work [11], regarding the same task, but in a Speaker Dependent (SD) environment, comparing results obtained in the three considered situations:

a) only the audio channel is active;
b) only the visual channel is active;
c) both audio and visual channel are simultaneously active,

the usefulness of visual parameters for improve speech recognition performance was successfully demonstrated. In this work, regarding a Speaker Independent (SI) environment, only the first and third situations in which only the audio channel or both audio and visual channels are simultaneously active were considered.

The network architecture which has been considered for the recognition was a fully connected recurrent feed-forward BP network with dynamic nodes positioned only in the hidden layer. The learning strategy was based on BPS algorithm [10], and only two supervision frames were chosen in order to speeding up the training time. The first one, focused on articulatory parameters, was positioned in the middle frame of the target plosive, the 'closure' zone, while the second, focused on acoustic parameters, was positioned in the penultimate frame, the 'burst' zone. A 20 ms delay, corresponding to 5 frames, was used for the hidden layer dynamic neurons. A 54(40+14)input ∗ 20(14+6)hidden ∗ (6)output RNN, as illustrated in Figure 1, was considered. Not all the connections were allowed from the input and the hidden layer, but only those concerning the two different modalities, which were thus maintained disjoint. Various parameter reduction schemes and various network structure alternatives were exploited but those described above represent the best choice in terms of learning speed and recognition performance.

RESULTS

Two different experimental setting were considered in which, among the 10 speakers, 8 speakers were randomly picked up in order to form the learning set while the remaining two were considered as the test set. The results for these two cases are illustrated in Table 1.

	E1	E2
Speaker 1	95.6	87.8
Speaker 2	72.2	66.7
Mean	83.9	77.8

Table 1. SI correct recognition rate in two experimental settings with 8 speaker for learning and 2 for testing.

After having observed that a particular speaker had a vary bad acquired audio signal, a third experiment was organized thus considering only 7 speakers for the learning set and two for the test set. Results regarding this case are summarized in Table 2.

In Table 3 the Speaker-Pooled (SP) mean correct recognition performance for all the three experimental settings is illustrated. In this case each speaker forming the learning set was also individually tested.

	E3
Speaker 1	95.6
Speaker 2	84.4
Mean	90.0

Table 2. SI correct recognition rate with the 9 speaker set (see text).

	E1	**E2**	**E3**
Mean	78.5	74.8	83.3

Table 3. SI mean correct recognition rate for the Speaker-Pooled (SP) case.

In order to test the power of the bimodal approach all the three experiments were repeated eliminating visual information thus retaining only the audio channel input. The 40 input $*$ 14 hidden $*$ 6 output RNN utilized in this case is exactly the audio subnet of the global net utilized in the bimodal environment as indicated in Fig. 1.

	E1	**E2**	**E3**
Mean	68.9	58.3	65.0

Table 4. SI mean correct recognition rate with only Audio information.

CONCLUSIONS

As indicated by a direct inspection of Tables 1-4, recognition performance significantly improves when both audio and visual channels are active. Looking at Table 3 referring to the speaker-pooled results a good generalization power can be associated with the chosen RNN given that SI results were surprisingly better than SP results.

REFERENCES

[1] P.L. Silsbee and A.C. Bovik (1993), "Medium-Vocabulary Audio-Visual Speech Recognition", *Proc. NATO ASI, New Advances and Trends in Speech Recognition & Coding*,pp. 13-16.

[2] D.W. Massaro (1987), "*Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry*", Lawrence Erlbaum Assoc., Hillsdale, New Jersey.

[3] B. Dodd & R. Campbell, Eds., (1987), "*Hearing by Eye: The Psychology of Lip-Reading*", Lawrence Erlbaum Assoc., Hillsdale, New Jersey.

[4] A. MacLeod and Q. Summerfield (1987), "Quantifying the contribution of vision to speech perception in noise", *British Journal of Audiology*, 21 pp. 131-141.

[5] C. Benoît (1992), "Bimodal Aspects of Speech Communication", personal communication.

[6] G. Ferrigno and A. Pedotti (1985), "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", *IEEE Transactions on Biomed. Eng.*, BME-32, pp. 943-950.

[7] S. Seneff (1988), "A joint synchrony/mean rate model of auditory speech processing", *Journal of Phonetics*, 16, 1988, pp. 55 76.

[8] P. Cosi(1992), "Ear Modelling for Speech Analysis and Recognition" (1992). In M. Cooke, S. Beet and M. Crawford (Eds.), *Visual Representation of Speech Signals*. John Wiley & Sons, pp.205-212.

[9] P. Cosi (1993), "SLAM: Segmentation and Labelling Automatic Module", *Proc. Eurospeech-93, Berlin*, 21-23 September, 1993, pp. 665-668.

[10] M. Gori, Y. Bengio and R. De Mori (1989), "BPS: A Learning Algorithm for Capturing the Dynamical Nature of Speech", *Proc. IEEE IJCNN89*, Washington, June 18 22, 1989, Vol. II, pp. 417 432.

[11] P. Cosi, E. Magno Caldognetto, K. Vagges, G.A. Mian, and M. Contolini (1994), "Bimodal Recognition Experiments with Recurrent Neural Networks", *Proc. IEEE ICASSP-94*, Adelaide, Australia, 19-22 April, 1994, Vol. 2, Session 20.8, pp. 553-556.