BIMODAL RECOGNITION EXPERIMENTS WITH RECURRENT NEURAL NETWORKS

P. Cosi, E. Magno Caldognetto, K. Vagges, G. A. Mian*, and M. Contolini*.

Centro di Studio per le Ricerche di Fonetica, C.N.R.
P.zza G. Salvemini 13, 35131 Padova (Italy).
* Universita' di Padova, Dipartimento di Elettronica ed Informatica,
Via Gradenigo, 35100 Padova, (Italy).

ABSTRACT

A bimodal automatic speech recognition system, using simultaneously auditory model and articulatory parameters, is described. Results given for various speaker dependent phonetic recognition experiments, regarding the Italian plosive class, show the usefulness of this approach especially in noisy conditions.

1. INTRODUCTION

Various studies of human speech perception have demonstrated that visual information plays an important role in the process of speech understanding [1], and, in particular, "*lip-reading*" seems to be one of the most important secondary information sources [2]. Moreover, even if the auditory modality definitely represents the most important flow of information for speech perception by normal or pathological subjects, the visual channel allows subjects to better understand speech when background noise strongly corrupts the audio channel [3]. Mohamadi and Benoît [4] 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). For these and other reasons audio-visual automatic speech recognition (ASR) systems can be conceived with the aim of improving speech recognition performance, mostly in noisy conditions [5-6].

2. METHOD

The system being described takes advantage of the jaw and lip reading capability, making use of a new system for automatic jaw and lips movement 3D analysis called ELITE [7], in conjunction with a joint synchrony/mean-rate auditory model of speech processing [8] which has shown great robustness in noisy condition [9]. 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 Recurrent Neural Network (RNN) utilized in the recognition phase [10].

The visual part of the system has adopted ELITE which is a fully automatic movement analyzer for 3D kinematic data acquisition. Previous phonetic studies have shown the usefulness of such a system in order to improve the knowledge of the human articulatory production mechanism [11], [12].

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 TVcameras 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 TVcameras 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 TVcameras. The collinearity equations [13] are iteratively linearized and solved at least squares after the acquisition of a known control object [14]. The 3D data coordinates are then used to evaluate the parameters described hereinafter.

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 [8] is applied producing 80 spectral-like parameters at 500 Hz frame rate. Only 40 parameters, out of the original 80, are used in the experiments being described in order to speed up the training time of the system. Due to the present complexity of the model, even if a quasi real-time implementation is already feasible [15], the auditory model is applied off-line. Input stimuli are segmented by SLAM, a recently developed semiautomatic segmentation and labelling tool [16] working on auditory model parameters.

Both audio and visual parameters, in a single or joint fashion, are used to train, by means of the Back Propagation for Sequences (BPS) [17] algorithm, an artificial Recurrent Neural Network (RNN) to recognize the input stimuli.

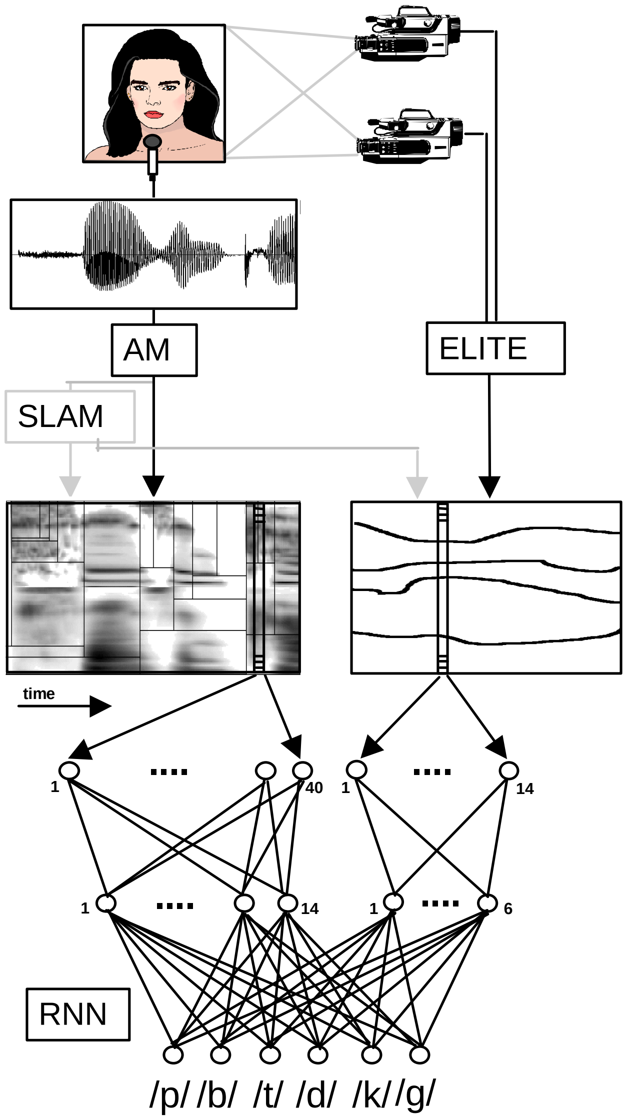

Figure 1. Block diagram of the overall bimodal recognition system.

3. EXPERIMENT

The input data consist of bysillabic symmetric /'VCV/ nonsense words, where C=/p,t,k,b,d,g/ and V=/a,i,u/, uttered by 2 male and 2 female 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.4), the corners of the mouth (n.3 and n.5), the jaw (n.6), and the lobe 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 4), together with the movements of the marker placed on the edges of the mouth (markers 3, 5), 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 4 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 4, the *lip opening width*, calculated as the distance between markers 3 and 5, and the *jaw opening* measured between the markers placed on the jaw and on the tip of the nose.

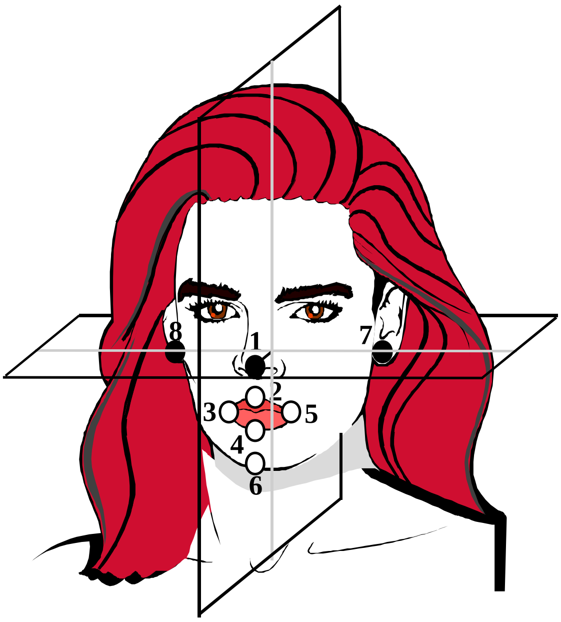

Figure 2. Position of the reflecting markers and of the reference plane. Identification numbers are indicated next to their corresponding markers. Marker dimension in the figure does not correspond to the real dimension (2mm) but is exaggerated for visualization purpose.

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. 4) associated with the sequence /'apa/.

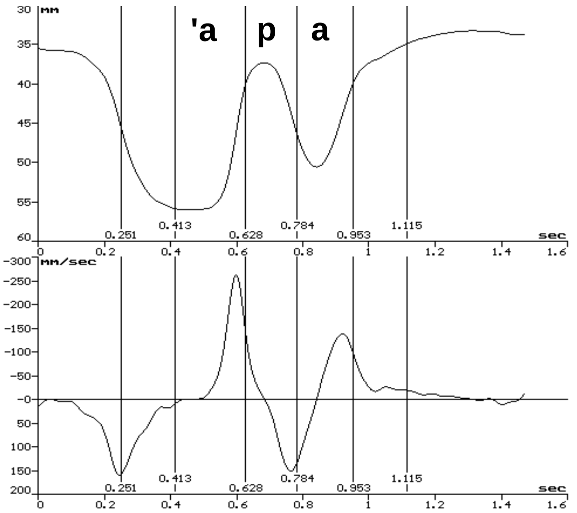

Figure 3. Time evolution of displacement and velocity of the marker placed on the lower lip (n.4), associated with the sequence /'apa/.

For a complete review regarding the application of the joint synchrony/mean-rate auditory model in speech analysis refer to [8] and [9]. As a simple example, spectral-like representations similar to that illustrated in Figure 1 at the output of the AM block can be obtained.

Three situations were considered in this study:

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

Moreover, for the two male speakers a very critical noisy condition of 0dB signal to noise ratio was also 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 [17] with only two supervision frames. The first one was positioned in the middle frame of the target plosive while the second was positioned in the penultimate frame. A 20 ms delay, corresponding to 5 frames, was used for the hidden layer dynamic neurons. In the first condition a 40x14x6 (input x hidden x output neurons) structure was considered while a 14x6x6 structure was used in the second one. In the third situation, when audio and visual channel are both simultaneously active, a (40+14)x(14+6)x6 structure, as illustrated in Figure 1, was considered. In this case, not all the connections are allowed from the input and the hidden layer, as in the previous conditions, 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

Table 1 summarizes the results obtained in all considered conditions for the four speakers in the clean speech case while Table 2 refers to the same experiments but in the noisy condition. In this case speech was corrupted by a white noise up to 0dB S/N ratio, which is a very hard condition for plosive recognition even for a human listener. It is immediately evident from Table 1, that articulatory parameters alone, give rise to quite poor performance in an open test case, but, it is worth mentioning that in the "close" case, when "place of articulation" (PLA in the Figure) classes were considered, grouping together:

- bilabial (/p/, /b/),
- dental (/t/, /d/),
- velar (/k/, /g/),

classification results significantly improved. Combining together acoustic (AM: Auditory Modelling) and articulatory (AR) parameters always improved the recognition rate in the clean case even if the acoustic information alone was rather satisfactory, given the very difficult task. As illustrated in Table 2, for the noisy case, the results show for both speakers considered, a significant improvement allowing the system to obtain similar performance to the clean case. The 97% mean recognition rate obtained in the clean condition with both acoustic and articulatory parameters and the identical 97% obtained in the 0dB S/N noisy condition represent a very attractive starting point for further development.

Speaker	AM	AR	AR (PLA)	AM+AR
MA (m)	83	67	100	94
LI (m)	78	61	89	94
PA (f)	78	67	100	100
AN (f)	72	72	100	100
mean	78	67	97	97

Table 1. Recognition performance in the clean condition.

Speaker	AM	AR	AR (PLA)	AM+AR
MA (m)	83	67	100	100
LI (m)	78	61	89	94
PA (f)	67	67	100	100
AN (f)	67	72	100	94
mean	74	67	97	97

Table 2. Recognition performance in the noisy condition (0dB S/N).

FUTURE TRENDS

More speakers will be analyzed in the near future but the experiments going on in different noisy conditions and with different modality seem to confirm our hypothesis and our first conclusions that, even if the visual channel alone is very poor to discriminate the input target stimuli, when both audio and visual channels are active recognition performance significantly improves especially when speech is highly degraded by environmental noise. The present study will obviously be extended to the speaker independent case. Being conscious that such a specialized hardware could not ever be included in any kind of present and future commercialized speech recognition system the aim of this work was only that of suggesting some articulatory parameters that can be of interest for recognition purpose and that could be similarly obtained by a direct inspection of the dynamic flow of the speaker image patterns taken by TVcameras synchronously with speech.

REFERENCES

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

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

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

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

[5] D.G. Stork, G. Wolff and E. Levine (1992), "Neural Network Lipreading System for Improved Speech Recognition", *Proceedings of Intl. Joint Conf. on Neural Networks*, pp. 285-295.

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

[7] 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.

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

[9] 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.

[10] P. Cosi, P. Frasconi, M. Gori and N. Griggio (1992), "Phonetic Recognition Experiments with Recurrent Neural Networks", *Proceedings International Conference on Spoken Language Processing* (ICSLP-92), Banff, Alberta, Canada, October 12-16, 1992, pp. 1335-1338.

[11] E. Magno Caldognetto, K. Vagges, G. Ferrigno, and G. Busà (1992). "Lip rounding coarticulation in Italian", *Proceedimgds of International Conference on Spoken Language Processing*, Banff 1992, Vol. 1, pp 61-64.

[12] E. Magno Caldognetto, K. Vagges, G. Ferrigno and Zmarich (1993). "Articulatory Dynamics of Lips in Italian /'VpV/ and /'VbV/ Sequences", *Proceedings. of Eurospeech-93*, Berlin, 21-23 September, 1993, Vol. 1, pp. 409-412.

[13] R.P. Wolf (1983), "*Elements of Photogrammetry*", Mc Graw-Hill Publisher, 1983.

[14] Borghese N.A., Ferrigno G., Pedotti A.(1988). "3D Movement Detection: a Hierarchical Approach, *Proceedings of the 1988 International Conference on Systems, Man and Cybernetics*, International Academic Publisher, 1988, pp.333-336.

[15] P. Cosi, L. Dellana, G.A. Mian and M. Omologo (1991), "Auditory Model Implementation on a DSP32C-Board", *Proc. GRETSI-91*, Juan Les Pins, 16-20 Sep 1991.

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

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