SPEAKER INDEPENDENT BIMODAL PHONETIC RECOGNITION EXPERIMENTS

P. Cosi[*]*, E. Magno Caldognetto*[*]*, F. Ferrero*[*]*, M. Dugatto*[**] *and K. Vagges*[*]

[*]Centro di Studio per le Ricerche di Fonetica (CNR)
Via Anghinoni, 10 - 35121 Padova (ITALY)
[**]Universita' di Padova, Dipartimento di Elettronica ed Informatica,
Via Gradenigo, 35100 Padova, (Italy).

ABSTRACT

A speaker independent *bimodal* phonetic classification experiment regarding the Italian plosive consonants is described. The phonetic classification scheme is based on a feed forward recurrent back-propagation neural network working on audio and visual information. The speech signal is processed by an auditory model producing spectral-like parameters, while the visual signal is processed by a specialized hardware, called ELITE, computing lip and jaw kinematics parameters.

1. INTRODUCTION

The idea of building new automatic speech recognition (ASR) systems able to use other sources of information than the acoustic signal such as those given by our visual channel, in order to improve, mostly in noisy conditions, current performance, is becoming more and more attractive within the scientific community, as underlined by the great attendance and success of the recent NATO Advanced Study Institute Workshop on "Speech Reading by Man and Machine: Models, Systems and Applications" [1].

This new trend is mainly due to the fact that various studies of human speech perception have shown that humans make use of various sources of information in order to recognize and understand speech with high accuracy, and they are able to visually classify classes of phonemes similarly produced by our articulators [2]. In other words, these studies have demonstrated that visual information plays an important role in the process of speech understanding [3], and, in particular, regarding "*speech-reading"*, that is the ability of tracking all facial expressions, "*lip-reading"* seems to be one of the most important secondary information sources [4]. 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 [5]. In fact, as Mohamadi and Benoît [6] reported, vision becomes essential when the noise highly degrades acoustic conditions (S/N 0dB).

Moreover, an impressive technological progress has been achieved in the field of image processing and probably all future personal computers will be equipped with a new generation of audio/visual sensors.

Similarly to other related studies appeared in the past [7], [8], [9], [10], the motivation of this work, is focused on the attempt of '*imitate*' human capabilities, while building new audio-visual ASR systems able of enhancing recognition performance, mostly in noisy conditions, but, differently by most of the other systems, the system being described in this work does not make use of a classical acoustic front-end processor, while uses instead a robust auditory speech processing scheme [11].

2. METHOD

The system being described, whose diagram is illustrated in Figure 1, is the same utilized in a previous speaker dependent experiment [12], and makes use of a specialized hardware for automatic jaw and lips movement 3D analysis called ELITE [13], [14], [15], in conjunction, as already underlined in the introduction, with a very robust feature extraction scheme in the acoustic domain, based on a well known joint synchrony/mean rate auditory model [11], which has shown great robustness in noisy condition [16], [17].

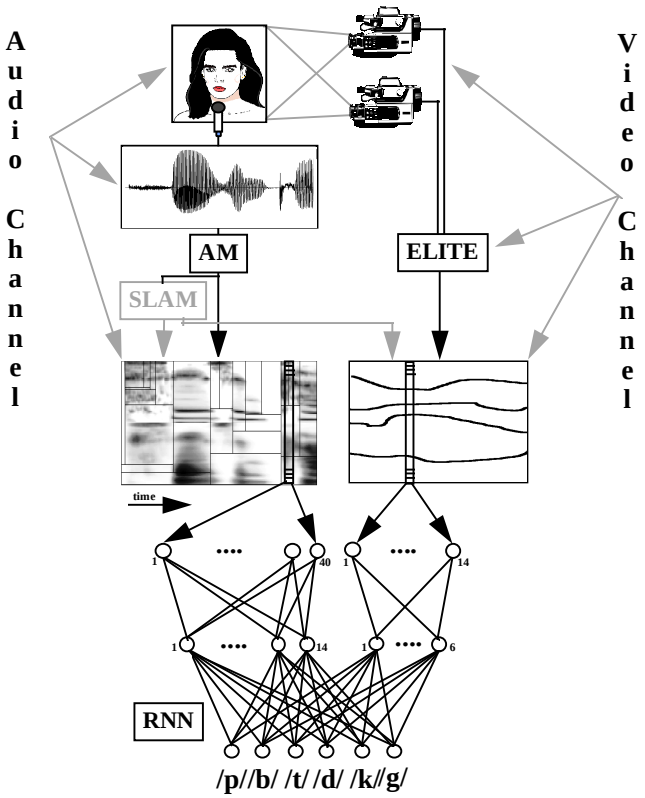

Figure 1. Block diagram of the overall bimodal recognition system.

The speech signal, acquired in synchrony with the articulatory data, is prefiltered and sampled at 16 kHz, and the auditory speech processing is applied producing a 80-dimensional spectral-like representation, such as that illustrated at the bottom of Figure 2 for the word /'iki/, at 500 Hz frame rate. Due to the present complexity of the model, even if a quasi real-time implementation is already feasible [18], the auditory processing is applied off-line. In the experiments being described, spectral-like parameters and frame rate have been reduced to 40 and 250 Hz respectively in order to speed up the system training time. Input stimuli were segmented, in the acoustic domain, by SLAM, a recently developed semi-automatic segmentation and labeling tool [19] working on auditory model parameters. In Figure 2 the segmentation proposed by SLAM for the stimulus /'iki/ is illustrated.

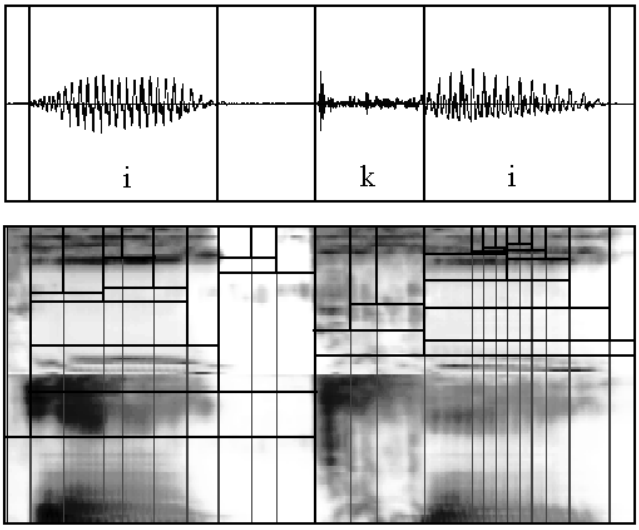

Figure 2. Envelope and synchrony auditory representation [11] relative to the nonsense word /'iki/. On the bottom it is evident the "*dendrogram*" structure built by SLAM [19] for segmentation, while on the top the proposed segmentation marks are illustrated.

The visual part of the system has adopted ELITE which is a fully automatic movement analyzer for 3D kinematics 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 collinearity equations [20] are iteratively linearized and solved at least squares after the acquisition of a known control object [21]. The 3D data coordinates are then used to evaluate the parameters described hereinafter.

Finally, as illustrated in Figure 1, both audio and visual parameters are used to train, by means of the Back Propagation for Sequences (BPS) algorithm [22], an artificial Recurrent Neural Network (RNN) to classify the input stimuli. Due to the different audio and visual frame rate, a 1:2.5 linear interpolation was adopted for visual parameters.

3. 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 producing the stimuli 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 nonsense 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. As illustrated in Figure 3, three reference points and five target points on the face of the subjects were considered.

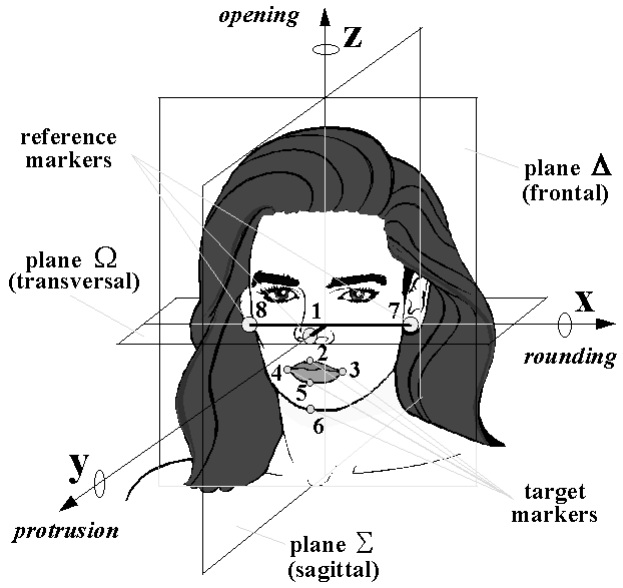

Figure 3. Position of the reflecting markers and of the reference planes. Identification numbers are indicated next to their corresponding markers. Marker dimension in the figure does not correspond to the real dimension (2mm) but is increased for visualization purpose.

In particular, 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 corners 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 transversal plane Ω depicted in Figure 3 and defined by the line crossing markers 7 and 8, placed on the ear lobes, and marker 1, placed on the tip of the nose. Similar distances with the frontal plane Δ 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, also listed in Table 1, were besides the upper and lower lip opening and closing movements (UL, LL), and the upper and lower lip protrusion (ULP, LLP), the lip opening height (LOH) calculated as the distance between markers 2 and 5, the lip opening width (LOW), calculated as the distance between markers 3 and 4, the jaw opening (JO), measured as the distance between the markers placed on the chin and on the tip of the nose, and the corresponding velocities.

code	meaning	definition
UL	upper lip vertical movement	d(m2,Ω)
LL	lower lip vertical movement	d(m5,Ω)
ULP	upper lip protrusion	d(m2,Δ)
LLP	lower lip protrusion	d(m5,Δ)
LOH	lip opening height	d(m2,m5)
LOW	lip opening width	d(m3,m4)
JO	jaw opening	d(m6,Ω)
ULv	∂UL/∂t	∂d(m2,Ω)/∂t
LLv	∂LL/∂t	∂d(m5,Ω)/∂t
ULPv	∂ULP/∂t	∂d(m2,Δ)/∂t
LLPv	∂LLP/∂t	∂d(m5,Δ)/∂t
LOHv	∂LOH/∂t	∂d(m2,m5)/∂t
LOWv	∂LOW/∂t	∂d(m3,m4)/∂t
JOv	∂JO/∂t	∂d(m6,Ω)/∂t

Table 1. Articulatory parameter definitions.

As an example of the articulatory parameters, Figure 4 shows the vertical, opening and closing, movement and the corresponding instantaneous velocity of the marker 5 placed on the lower lip (LL, LLv) associated with the sequence /'apa/.

Finally a recurrent feed-forward BP network with dynamic nodes, such as that illustrated in Figure 5, positioned only in the hidden layer was used for the classification experiment. As illustrated in Figure 1, 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, while, on the contrary, output and hidden layers were fully connected.

In order to reduce the training time, the learning strategy was based on BPS algorithm [22] 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 particular a $(40+14)^{*}(14+6)^{*}6$ structure, as it is shown at the bottom of Figure 1, was considered. 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.

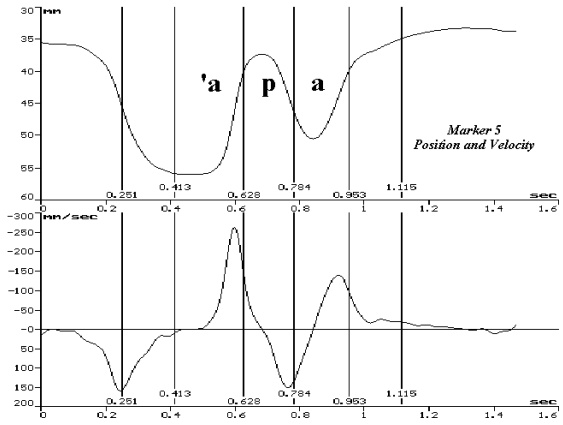

Figure 4. Time evolution of displacement and velocity of the marker placed on the lower lip (n.5), associated with the sequence /'apa/.

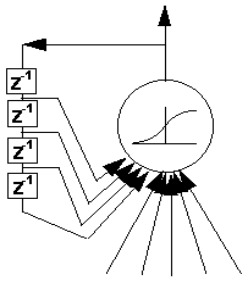

Figure 5. Typical dynamic node structure with 4 delay blocks.

RESULTS

Given the small number of speakers, in order to increase the statistic relevance of the data, the same classification experiment was repeated 10 times following the so-called “jack-knife” technique, where 9 speakers were considered for learning and 1 testing. The results reported in Table 1 indicate a rather good 71% correct classification performance in the “open” case, when all the plosives were separately considered. In the “close” case, i. e. when "PLace of Articulation" (PLA in the Table) classes were considered grouping together bilabial (/p/, /b/), dental (/t/, /d/), and velar (/k/, /g/),as occurred also in a previous Speaker Dependent (SD) experiment on the same phonetic class [12], classification results significantly improved up to 77% correct classification.

talker	% correct	% correct PLA
talker 1	84.4	87.8
talker 2	83.3	83.3
talker 3	93.3	93.3
talker 4	64.4	71.1
talker 5	47.8	56.7
talker 6	53.3	64.4
talker 7	75.6	76.7
talker 8	60.0	76.7
talker 9	62.2	71.1
talker 10	86.7	91.1
mean	71.1	77.22

Table 4. Speaker Independent correct recognition rate (%) for the 10 repeated trials ("jack knife" technique).

CONCLUSIONS

The results given for the SI plosive classification experiment were rather good given the quite difficult task of classifying these consonants using only two supervision points. Given the difficulty to include a specialized hardware like the one described in this work in any kind of present commercialized speech recognition system, the aim of this work was simply to suggest some articulatory parameters that can be of interest for classification purpose and that can be also obtained by a direct inspection of the dynamic flow of the speaker image patterns taken by TV cameras synchronously with speech.

REFERENCES

[1] D. G. Stork and M. Henneke (eds.). "Speech Reading by Man and Machine: Models, Systems and Applications", NATO ASI Series, Series F: Computer and System Sciences, *Proc. Nato Advanced Study Institute*, Château de Bonas, France, August 28- September 8, 1995, (to be published).

[2] E. Magno Caldognetto, K. Vagges, and F. Ferrero. "Un test di confusione fra le consonanti dell'italiano: primi risultati", *Atti del Seminario "La percezione del linguaggio"*, Accademia della Crusca, Firenze, 17-20 Dicembre, 1980, pp. 123-179.

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

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

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

[6] Benoit C., Lallouache T., Mohamadi T., and Abry C.. "A Set of French Visemes for Visual Speech Synthesis", in Bailly G., Benoit C., and Sawallis T.R. (Eds.), *Talking machines: Theories, Models, and Designs*, North-Holland, Amsterdam, 485-504.

[7] E.D. Petajan. "*Automatic Lipreading to Enhance Speech Recognition*", PhD Thesis, Univ. of Illinois at Urbana-Champaign, 1984.

[8] D. G. Stork, G. Wolff and E. Levine. "Neural Network Lipreading System for Improved Speech Recognition", *Proc. IEEE IJCNN-92, 1992, pp.* 285-295.

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

[10] A. Adjoudani and C. Benoit. "Audio-Visual Speech Recognition Compared Across Two Architectures", *Proc. Eurospeech-95*, Madrid, Spain, 18-21 September 1995, Vol. 2., pp. 1563-1566.

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

[12] P. Cosi, E. Magno Caldognetto, K. Vagges, G.A. Mian, and M. Contolini. "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.

[13] G. Ferrigno and A. Pedotti. "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", *IEEE Trans. on Biomedical Engineering*. BME-32:943-950, 1985.

[14] E. Magno Caldognetto, K. Vagges, G. Ferrigno, and G. Busà. "Lip Rounding Coarticulation in Italian", *Proc. ICSLP-92*, Banff, Canada, 1992, Vol. 1, pp. 61-64.

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

[16] C.R. Jankowski Jr., H-D. H. Vo and R.P. Lippmann. "A Comparison of Signal Processing Front Ends for Automatic Word Recognition", *IEEE Trans. on Speech and Audio Processing*, Vol. 3, N. 4, July, 1995, pp. 286-293

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

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

[19] P. Cosi. "SLAM: Segmentation and Labeling Automatic Module", *Proc. Eurospeech-93*, Berlin, Germany, September, 21-23, 1993, pp. 665-668.

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

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

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

Phone: (+39) 49 8284050, FAX: (+39) 49 8754556
EMail: COSI@CSRF00.CSRF.PD.CNR.IT