Phonetic Recognition by Recurrent Neural Networks Working on Audio and Visual Information

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

Centro di Studio per le Ricerche di Fonetica - C.N.R.
Department of Linguistics, University of Padua
Via G. Anghinoni, 10 - 35121 Padova, Italy

Abstract

A phonetic classification scheme based on a feed forward recurrent back-propagation neural network working on audio and visual information is described. The speech signal is processed by an auditory model producing spectral-like parameters, while the visual signal is processed by a specialised hardware, called ELITE, computing lip and jaw kinematics parameters. Some results will be given for various speaker dependent and independent phonetic recognition experiments regarding the Italian plosive consonants.

[*] Email: cosi@csrf00.csrf.pd.cnr.it

1. Introduction

Humans make use of various sources of information in order to recognise and understand speech with high accuracy. Various studies of human speech perception have demonstrated that visual information plays an important role in the speech understanding process (Massaro, 1987). Speaking of "speechreading", that is the ability of tracking all facial expressions, "lip-reading" seems to be one of the most relevant secondary information sources (Dodd & Campbell, 1987) for understanding the communication message. 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 (MacLeod & Summerfield, 1987). In fact, as Mohamadi and Benoît (Mohamadi & Benoît, 1992) reported, vision becomes essential when the noise highly degrades acoustic conditions (S/N ≤ 0dB). Various studies appeared in the literature showing how humans are able to visually classify classes of phonemes similarly produced by our articulators (as for Italian, refer to Magno Caldognetto et al. 1980). 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. Thus, the idea of building new automatic speech recognisers able to use other sources of information than the acoustic signal, such as those given by our visual channel, is becoming more and more attractive within the scientific community, as underlined by the great attendance and success of the recent Workshop on “Speech Reading by Man and Machine: Models, Systems and Applications” organised by the NATO Advanced Study Institute (Stork & Henneke, 1995). The motivation of this work, which is essentially the same of all other related studies appeared in the past, from the first paper of E. Petajan (Petajan E., 1984) to the more recent works (Stork et al. 1992, Silsbee & Allen, 1993, Adjoudani & Benoit, 1995), is focused on the attempt of building a new audio-visual automatic speech recognition (ASR) systems able of enhancing recognition performance, mostly in noisy conditions. Differently by most of the other systems, the system being described in this work, instead of using a classical acoustic front-end processor, makes use of a well known joint synchrony/mean rate auditory model (Seneff 1988), in order to use very robust features in the acoustic domain, as stated by

Jankowski et al. 1995, and still verify the usefulness of visual information..

2. Experiment

The system being described, whose diagram is illustrated in Figure 1, makes use of a new system for automatic jaw and lips movement 3D analysis called ELITE (Ferrigno & Pedotti 1985, E. Magno Caldognetto et al. 1992, 1993), in conjunction, as already underlined in the introduction, with an auditory model of speech processing (Seneff 1988) which has shown great robustness in noisy condition (Cosi 1992).

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 (Seneff 1988) is applied producing 80 spectral-like parameters at 500 Hz frame rate. Due to the present complexity of the model, even if a quasi real-time implementation is already feasible (Cosi et al. 1991), the auditory model 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 labelling tool (Cosi 1993) working on auditory model parameters.

The visual part of the system has adopted ELITE which is a fully automatic movement analyser 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, realised 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 characterised 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 recognised, the cross-correlation algorithm allows the computation of the weighted centre of mass increasing the accuracy of the system up to 0.1mm on 28cm of field of view. The coordinates of the recognised 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 (Wolf 1983) are iteratively linearised and solved at least squares after the acquisition of a known control object (Borghese et al.1988). 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) algorithm (Gori et al. 1989), 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.

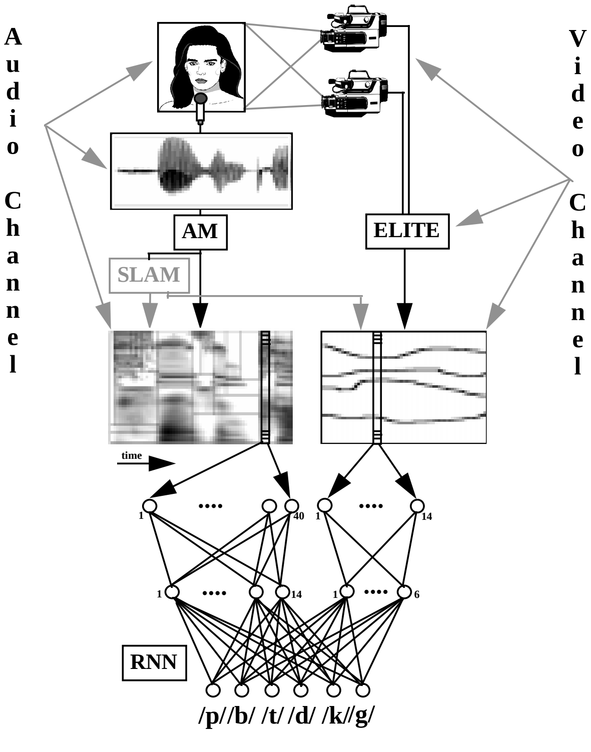


Figure 1. Block diagram of the bimodal recognition system.

In all the experiments described in the following sections the input data consist of disyllabic symmetric /'VCV/ nonsense words, where C=/p,t,k,b,d,g/ and V=/a,i,u/. 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 2, three reference points and five target points on the face of the subjects were considered.

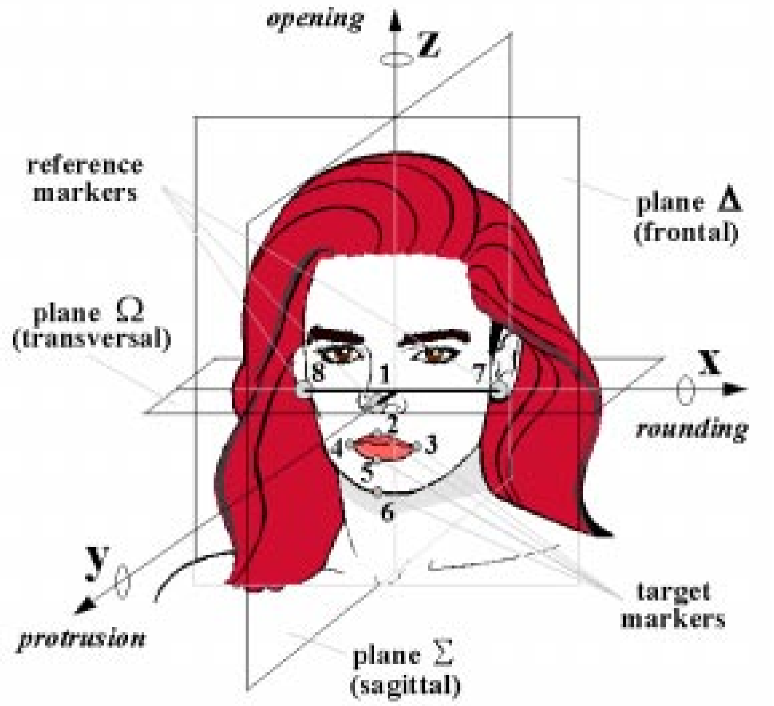


Figure 2. Position of the reflecting markers and of the reference planes. Identification numbers are indicated next to their corresponding markers. Marker dimension in the figure does not correspond to the real dimension (2mm) but is increased for visualisation purpose.

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 analysed, 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 2 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 opening and closing movement	d(m2,Ω)
LL	lower lip opening and closing 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 parameters.

As an example of the articulatory parameters, Figure 3 shows the 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/.

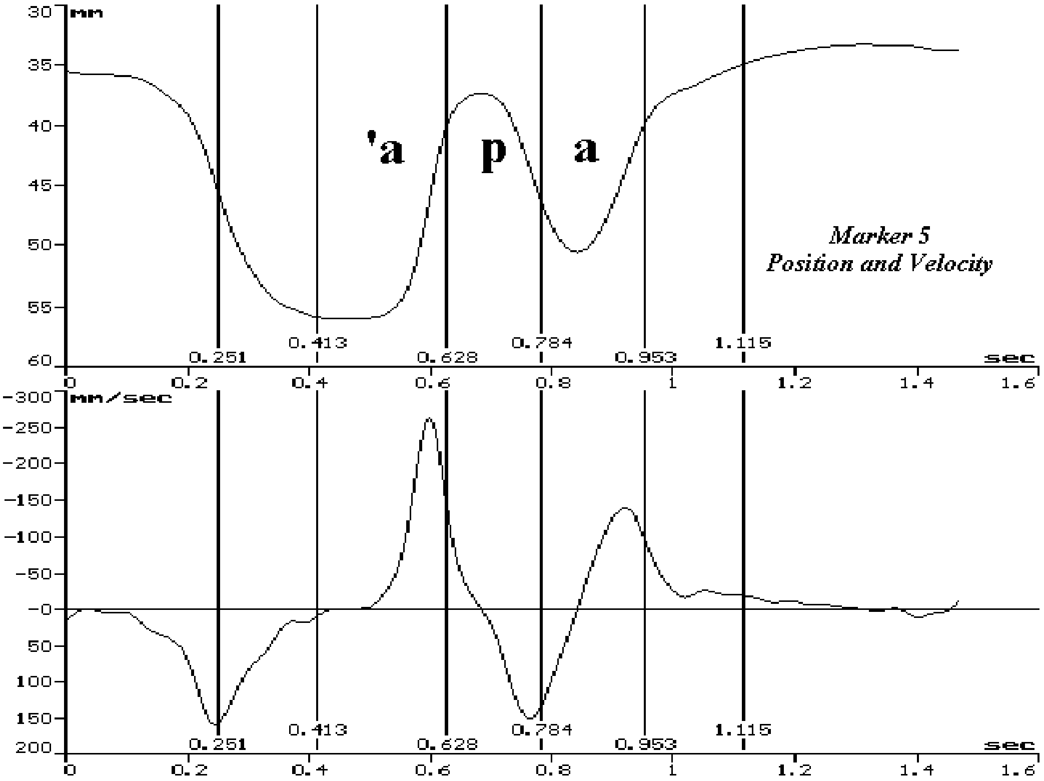


Figure 3. Time evolution of displacement and velocity of the marker placed on the lower lip (n.5), associated with the sequence /’apa/.

3. Speaker dependent case

For this experiment (Cosi et al. 1994), 2 male and 2 female speakers in three different experimental settings were considered:

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

Moreover a critical noisy condition of 0dB signal to noise ratio was tested. The network architecture which has been considered for the recognition was a fully or partially connected recurrent feed-forward BP network with dynamic nodes positioned only in the hidden layer.

The learning strategy was based on BPS algorithm and, as illustrated in Figure 4, only two supervision frames were chosen in order to speed up the training procedure time. The first one, focused on articulatory parameters, was positioned in the middle frame of the target plosive (the 'closure' zone), as defined by the auditory-based SLAM segmentation procedure, while the second, focused on acoustic parameters, was positioned in the penultimate frame (the 'burst' zone). During the testing phase all the frames of the target consonant were considered but only in the second supervision point, in other words at the end of the stimulus, the output was analysed.

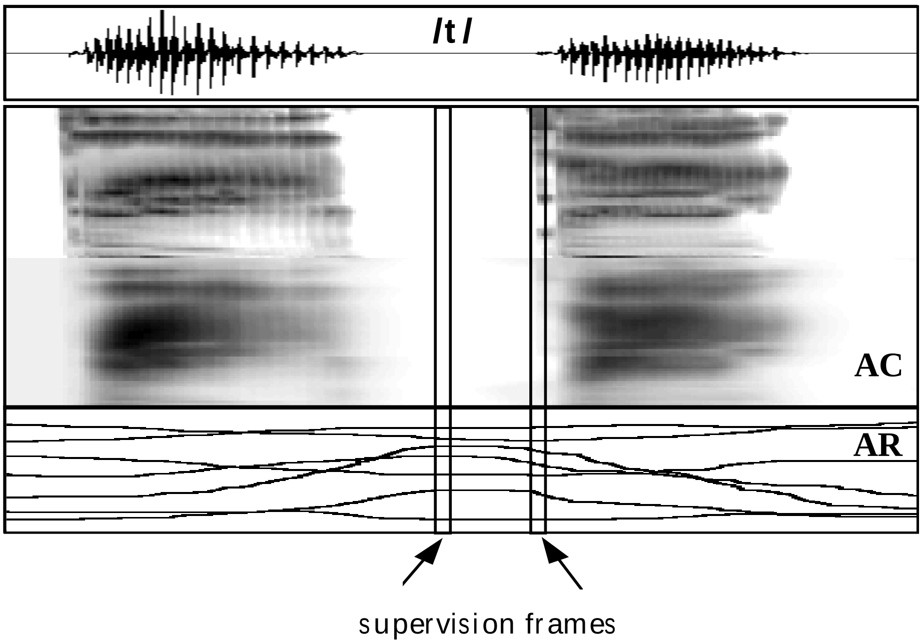


Figure 4. Two target points supervision strategy for the sequence /'ata/. Signal waveform, ACoustic and ARticulatory representations are illustrated from the top.

A 20 ms delay, corresponding to 5 frames, was used for the dynamic neurons belonging to the hidden layer. In the first (a) condition a 40(input)*14(hidden)*6(output) RNN structure was considered, while a 14(input)*6(hidden)*6(output) structure was used in the second (b) condition. In the third (c) condition, when audio and visual channel are both active, a 54[40+14](input)*20[14+6](hidden)*6(output) structure was adopted. In this case, not all the connections were 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 and graphically summarised in Figure 5, represent the best choice in terms of learning speed and recognition performance.

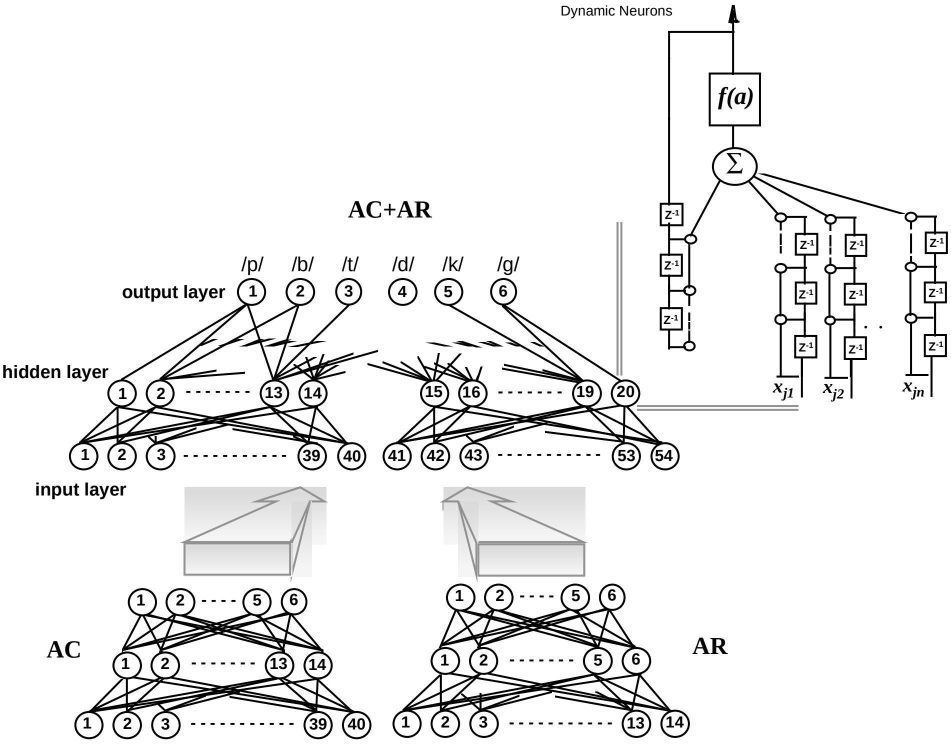


Figure 5. Network structures in the three different experimental settings (see also text).

Table 2 and 3 summarise the results obtained for the four speakers respectively for the clean and noisy conditions. In the noisy case the speech signal was corrupted by a white noise with 0dB S/N ratio, which is a very hard condition for plosive recognition, even for a human listener. For each speaker 5 experiments were executed using 4 repetitions of the input stimuli for learning and one for testing. Thus the results shown in the Tables 2 and 3 represent the means of the 5 experiments. Looking at the Tables, it is immediately evident that articulatory parameters alone give rise to quite poor performance in an open test case. On the contrary, in the *close* case, when *place of articulation* (PLA in Table 2) classes were considered, grouping together bilabial (/p/, /b/), dental (/t/, /d/), and velar (/k/, /g/) consonants, the classification results significantly improved (99%). Combining together acoustic (AC) and articulatory

(AR) parameters always improved the recognition rate in the clean case even if the acoustic information alone was rather satisfactory. As for the noisy case, the results show, for all the speakers, a significant improvement using both AC and AR parameters than using AC parameters alone, allowing the system to obtain similar performance (96%) to the clean case (98%).

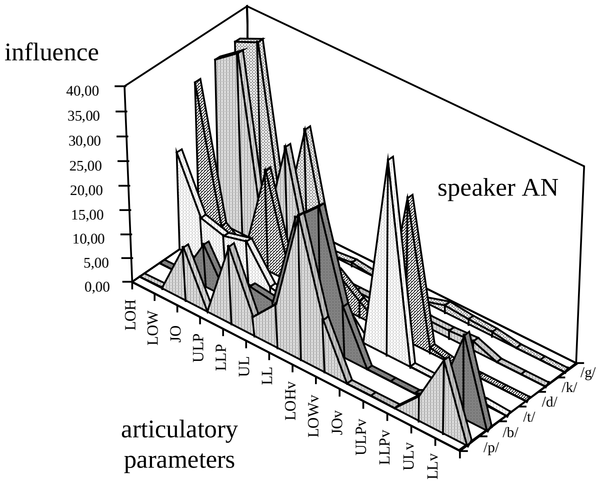
talker	**AC**	**AR**	**AR(PLA)**	**AC+AR**
MA(m)	83	67	100	100
LI(m)	78	61	97	97
PA(f)	78	67	98	96
AN(f)	72	72	100	98
mean	78	67	99	98

Table 2. Speaker Dependent correct recognition rate (%) in the clean condition. AC: acoustic parameters, AR: articulatory parameters,(PLA): Place of Articulation (see text).

talker	**AC**	**AC+AR**
MA(m)	83	100
LI(m)	78	94
PA(f)	67	95
AN(f)	67	94
mean	74	96

Table 3. Speaker Dependent correct recognition rate (%) in the noisy condition. AC: acoustic parameters, AR: articulatory parameters.

In order to *qualitatively* explain the network ability to well discriminate among the different classes of articulation (99 %), the influence of input articulatory parameters on the correct outputs was measured. The data shown in Figure 6 refers to the speaker AN and were obtained considering the learning set as the test set in order to have a big number of correct classifications. For each corrected classified plosive, going backward to the hidden layer, the hidden nodes which *positively* activate the correspondent output node were considered. Successively, going backward to the input layer, the input nodes which positively activate the just evidenciated hidden nodes were isolated and a sort of *influence measure* was computed for each input node by considering the output values of the nodes crossed by positive activating links. Various interesting qualitative deductions can be drawn. In particular it can be observed that a similar activation pattern represents the same plosive class. For example, a high influence measure of LOHv (lip opening height velocity) and also a low influence measure of ULP (upper lip protrusion) well differentiate bilabial consonants /p/ and /b/ from the other two classes. Similar plots, obtained for other speakers, show a similar tendency thus justifying the results illustrated in Table 2. These observations obviously need further investigation which will be completed in the future. In particular, a more complete statistical description of articulatory data will be computed, so as to justify the hypothesised ability of the chosen RNN to identify the most valuable and reliable parameters for the PLA class discrimination.



par/out	*/P/*	*/B/*	*/T/*	*/D/*	*/K/*	*/G/*
LOH	0,00	0,00	20,18	31,74	33,58	34,56
LOW	0,13	0,00	9,02	4,46	37,12	36,66
JO	10,95	8,68	7,61	1,01	1,43	1,18
ULP	0,00	0,00	8,86	20,73	22,72	23,54
LLP	16,00	4,77	1,78	1,49	0,13	0,09
UL	3,85	4,14	0,81	0,05	0,05	0,65
LL	9,03	23,92	10,04	6,97	0,00	0,00
LOHv	29,44	28,48	5,35	3,26	0,00	0,00
LOWv	10,73	10,49	0,00	0,00	0,00	0,00
JOv	0,00	0,00	39,59	29,25	1,56	1,49
ULPv	0,00	0,00	0,06	1,03	1,56	1,00
LLPv	0,00	0,00	0,00	0,00	1,85	0,82
ULv	5,06	2,44	0,00	0,00	0,00	0,00
LLv	15,05	17,08	0,33	0,00	0,00	0,00

Figure 6. Influence measure of input articulatory parameters for the recognition of plosives relatively to the speaker AN (see text).

4. Speaker independent case

In the Speaker Independent (SI) case 10 male talkers were considered. 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. For this experiment: both audio and visual channel were considered, the speech material was recorded only in a clean condition, and, for a sake of simplicity, the network architecture and the learning strategy were identical to that chosen for the SD case, even if a more complex structure could be better in this situation. In other words, the same 54[40+14](input)*20[14+6](hidden)*6(output) network structure and the same learning strategy, based again on BPS algorithm with only two supervision frames, in order to speed up the training procedure time (see Section 3), were adopted. The results shown in Table 4 refer to the condition c), previously described, when both AC and AR parameters were considered as input to the classification network. These results 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 Table 4) classes were considered grouping together bilabial (/p/, /b/), dental (/t/, /d/), and velar (/k/, /g/), as executed in the previous SD experiment (see Table2), 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). Both AC and AR parameters were considered. The second column refers to the "open" case, when all the plosives were separately considered, while the third one refers to the "close" case, when "PLace of Articulation" (PLA) classes were considered grouping together bilabial (/p/, /b/), dental (/t/, /d/), and velar (/k/, /g/) consonants.

5. Conclusions

As indicated by a direct inspection of Tables 2-3 for the SD experiment, recognition performance significantly improves when both audio and visual channels are active. Looking at Tables 4 referring to the SI case, a good generalisation power can be associated with the chosen RNN given that SI results were rather good principally considering the quite difficult task of recognising plosives using only two supervision points. Given the difficulty to include a specialised hardware like the one described in this work in any kind of present commercialised speech recognition system, the aim of this work was to suggest some articulatory parameters that can be of interest for recognition 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

A. Adjoudani & C. Benoit (1995). Audio-Visual Speech Recognition Compared Across Two Architectures. *Proc. of Eurospeech-95*, 18-21 Sept. 1995, Madrid, Spain, Vol. 2., 1563-1566.

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

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

P. Cosi, L. Dellana, G.A. Mian & M. Omologo (1991). Auditory Model Implementation on a DSP32C Board. *Proc. GRETSI-91*. Juan Les Pins, France. September, 16-20, 1991.

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

P.Cosi, E. Magno Caldognetto, K. Vagges, G.A. Mian, & M. Contolini (1994). Bimodal Recognition Experiments with Recurrent Neural Networks. *Proc. of IEEE International Joint Conference on Acoustics Speech and Signal Processing, ICASSP-94*. Adelaide, Australia. April 19-22. paper 20.8.

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

G. Ferrigno & A. Pedotti (1985). ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing. *IEEE Transactions on Biomedical Engineering*. BME-32:943-950.

M. Gori, Y. Bengio & R. De Mori (1989). BPS: A Learning Algorithm for Capturing the Dynamical Nature of Speech. *Proc.*

IEEE International Joint Conference on Neural Networks, IJCNN-89. Washington DC, USA. June 18-22, 1989. II:417:432.

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

A. MacLeod & Q, Summerfield (1987). Quantifying the Contribution of Vision to Speech Perception in Noise. *British Journal of Audiology*. 21:131-141.

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

E. Magno Caldognetto, K. Vagges, G. Ferrigno, & G. Busà (1992). Lip Rounding Coarticulation in Italian. *Proc. International Conference on Spoken Language Processing, ICSLP-92*. Banff, Canada. 1992. 1:61-64.

E. Magno Caldognetto, K. Vagges, G. Ferrigno, & C. Zmarich (1993). Articulatory Dynamics of Lips in Italian /'VpV/ and /'VbV/ Sequences. *Proc. Eurospeech-93*. Berlin, Germany. September 21-23, 1993. 1:409-412.

D.W Massaro (1987). S*peech Perception by Ear and Eye: a Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

T. Mohamadi & C. Benoît (1992). Apport de la vision du locuteur à l'intelligibilité de la parole bruitée en français. *Bulletin de la Communication Parlée*. 2:31-41.

E.D. Petajan (1984). Automatic Lipreading to Enhance Speech Recognition, *PhD Thesis*, Univ. of Illinois at Urbana-Champaign.

S. Seneff (1988). A joint synchrony/mean rate model of auditory speech processing. *Journal of Phonetics*. 16:55-76.

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

D. G. Stork, G. Wolff & E. Levine (1992). Neural Network Lipreading System for Improved Speech Recognition, *Proc. of IEEE International Joint Conference on Neural Networks, IJCNN-92.* 285-295.

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

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