SPEAKER INDEPENDENT PHONETIC RECOGNITION USING AUDITORY MODELLING AND RECURRENT NEURAL NETWORKS

P. Cosi, 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).
EMail: cosi@csrf.pd.cnr.it
*Università di Padova, Dipartimento di Elettronica ed Informatica, Padova (Italy).

ABSTRACT

Two speaker independent speech recognition experiments, regarding the automatic discrimination of the Italian alphabet **I-set** and **E-set**, two very difficult Italian phonetic classes, will be described. The speech signal is analyzed by a recently developed joint synchrony/mean-rate auditory processing scheme and a fully-connected feed-forward recurrent BP network was used for the classification stage. The achieved speaker independent mean recognition rate was 65%, for the I-set and 88% for the E-set showing rather satisfactory results given the difficulty of both tasks.

1. INTRODUCTION

Both static and dynamic networks have been proposed in speech technology, and especially in speech recognition, as opposed to more classic statistical approaches. Experimental results show that neural networks represent an effective alternative to classical pattern recognition methods in several applications. The Multi-Layered Neural Networks (MLN) trained with Back-Propagation (BP) are probably the most used as static networks.

Instead of transposing static network techniques to the continuous case, the direct use of dynamic network architectures seem to be the best solution to tackle the problem of continuous speech recognition. For this reason a simple Dynamic Multi-Layered Neural Network (DMLNN) with local feedback connections, trained by Back Propagation for Sequences (BPS) [1], was used, in this study, in order to discriminate input speech stimuli.

A physiologically-based auditory speech processing [2] was chosen as a front-end instead of a classical approach, like FFT, LPC or CEPSTRUM based filter bank.

2. METHOD

A vector of 40+40 spectral-like parameters, representing the "mean rate" and the "synchronous" response of auditory neurons, produced by a joint synchrony/mean-rate auditory model [2], was used and presented to the network at a certain frame-rate, both during the learning and the testing phase. Input stimuli were

semi-automatically segmented by the use of SLAM [3], a semi-automatic segmentation and labelling tool working on auditory model parameters. Advantages of using an auditory model for automatic speech segmentation have been shown especially in adverse conditions [4]. During the learning supervision, not only the knowledge of the stimulus identity was available, but also its fine segmentation characteristics. A simple DMLNN, trained by BPS [1], was used for recognition. Supervision was executed without considering a static input, thus the inherent limitation of most dynamic neural networks currently used in speech technology, was overcome. Instead of waiting for a fixed point a learning algorithm was used in which the output supervision was done during the evolution of the activations. The learning environment was defined by a sequence of frames representing the natural time evolution of speech signals. The dynamic model considered was discrete instead of continuous and its transitions occurred when a new frame was applied at the input. The class of DMLNNs utilized in this experiment was a simple one, in which the dynamic neurons, with local feedback connections to themselves, have only incoming connections from the input layer. Learning was organized within an isolated word recognition framework, and the network should output the right answer after presenting each unknown stimulus.

3. EXPERIMENT

The two experiments described in this paper regard the automatic speaker independent recognition of the following two Italian phonetic classes:

a) I-set: /bi/, /tSi/, /di/, /dZi/, /i/, /pi/, /ti/, /vi/ plus other two "out of alphabet" stimuli /Li/, /si/

b) E-set: /'Effe/, /'Elle/, /'Emme/, /'Enne/, /'Erre/, /'Esse/,

(see SAMPA Phonetic Alphabet [5]).

Speech data-base is made up of 7 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. A total of 350 stimuli for the I-set and 210 stimuli for the E-set were thus available for training and testing the two recurrent neural networks. Circularly one speaker (50 stimuli for the I-set and 30 stimuli for the E-set) was tested, using the remaining 6 speakers for learning (300 stimuli for the I-set and 180 stimuli for the E-set).

The dynamic networks utilized in the two experiment were slightly different, their structure resulting from a trial and error procedure. Figure 1 shows the two different architectures utilized. Both networks had a MLN architecture in which both static and dynamic neurons cooperate. In particular, in both cases, a very simple DMLNN structure was used, in which dynamic neurons, with local feedback connections to themselves, had only incoming connections from the input layer. Frame rate was set to 2ms for the I-set experiment and to 8ms for the E-set experiment. The delay value was set to 4 frames for both experiments. The two DMLNN architectures are illustrated in Fig. 2. 80 static neurons were considered at the input level. They received, frame by frame, the output of the auditory front-end. 20 dynamic neurons with a 4-delay dynamics were considered at the hidden layer and 10 static neurons, one for each target phonetic stimulus, were considered at the

output level, for the I-set experiment. For the E-set, since the number of the output neurons was reduced to 6, the number of the hidden neurons was reduced to 8. Learning supervision time was forced only at the last frame of the target stimuli.

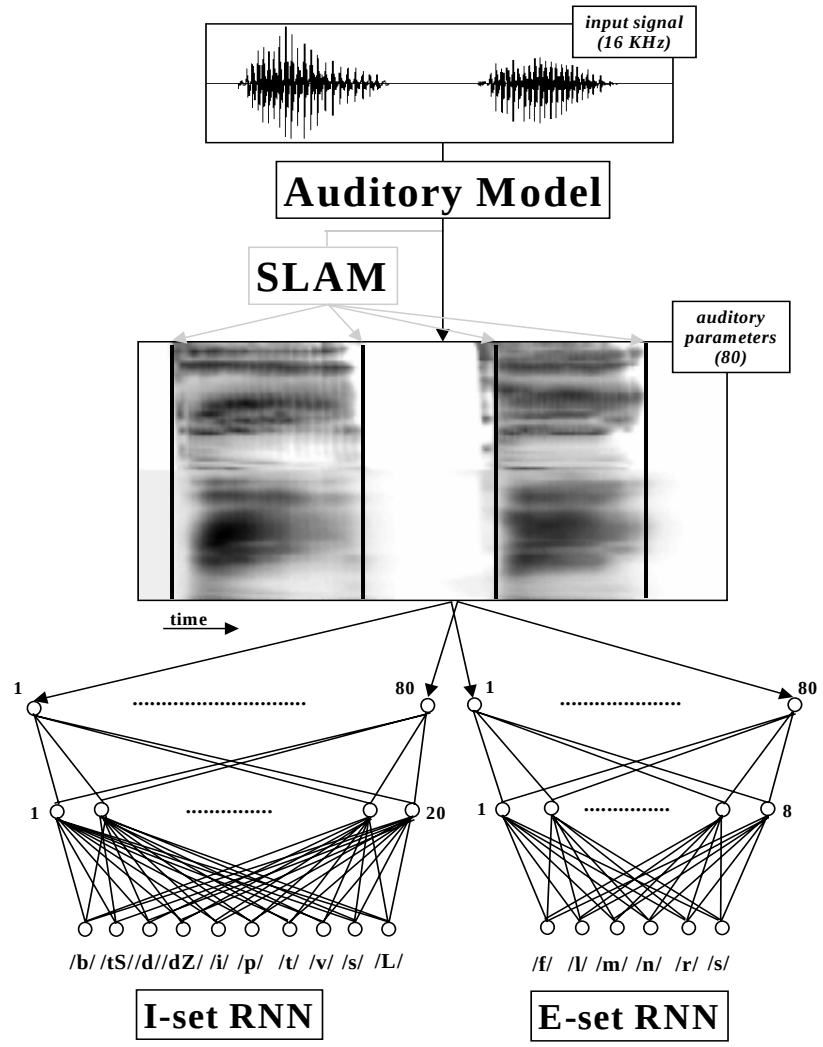

Figure 1. Block diagram of the recognition system. Both network architectures are illustrated.

4. RESULTS

Given the low number of speakers, results are given mediating 6 different experimental sets. Circularly one speaker was used as the test speaker, while the

remaining 6 were used for training the DMLNN. In Table 1, all the correct recognition rates and the global mean rate (MM) are illustrated for the I-set and the E-set experiments. Analyzing the results of the first experiment it can be observed that two speakers, SR and BC, were particularly difficult to recognize, achieving only 52% correct recognition rate, while one subject had low rates in the second experiment achieving only 70% correct recognition rate. For the I-set, the best speaker achieved 80% correct recognition rate, which is, given this particular difficult task, a very promising result., while for the E-set the best speaker achieved a 97% correct recognition rate.

Speaker	I-set Error Rate	E-set Error Rate
MM	22	7
GF	32	30
PT	36	13
SR	48	10
BC	48	10
EP	36	10
MR	27	3
MM	22	12

Table 1. I-set and E-set recognition performance (error rate, %).

More speakers will be analyzed in order to confirm our preliminary results and more phonetic classes will be studied in order to build a complete neural speaker independent phoneme classifier. Other modalities, such as some articulatory cues, will be included in the near future in order to improve speech recognition performance, especially in difficult noisy conditions.

REFERENCES

[1] 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.

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

[3] P. Cosi (1993), "SLAM: Segmentation and Labelling Automatic Module", *Proceedings of Eurospeech-93*, Berlin, 21-23 September, 1993, pp665-668.

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

[5] A.J. Fourcin, G. Harland, W. Barry and W. Hazan eds. (1989), "*Speech Input and Output Assessment, Multilingual Methods and Standards"*, Ellis Horwood Books in Information Technology, 1989.