AUTOMATIC RECOGNITION OF ITALIAN I-SET BY RECURRENT NEURAL NETWORKS

P. Cosi[*], P. Frasconi[**], M. Gori[**] and N. Griggio[***]

[*]Centro di Studio per le Ricerche di Fonetica,
Consiglio Nazionale delle Ricerche,
P.zza G. Salvemini, 13 - 35131 Padova (Italy).

[**]Dipartimento di Sistemi e Informatica,
Università degli Studi di Firenze,
Via di S. Marta, 3 - 50139 Firenze (Italy).

[***]Dipartimento di Elettronica ed Informatica,
Università degli Studi di Padova,
Via G. Gradenigo, 6 - 35100 Padova (Italy).

ABSTRACT

In order to prove the potential power of "learning by examples" paradigm for problems of Automatic Speech Recognition, an experiment is described, regarding an extremely difficult Italian phonetic recognition problem:

the automatic discrimination of the so called Italian **i-set**:

/bi/, /tSi/, /di/, /dZi/, /i/, /pi/, /ti/, /vi/
plus
other two i-like stimuli **/Li/, /si/.**

Auditory Modeling is used as front-end digital signal processing. Semi-automatic Multi-Level segmentation is applied to input speech stimuli. Recurrent Neural Networks trained by Extended Back Propagation for Sequences constitute the global recognition framework.. The achieved speaker independent mean recognition rate is around 65% which, given the effective difficulty of the present task, can be considered quite acceptable and promising.

INTRODUCTION

Several neural network models have been recently investigated by researchers for dealing with signal processing and particularly with automatic speech recognition. Both static and dynamic networks have been proposed and experimental results already 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 [1]. A dynamical behavior has been differently added to MLNs thanks to various techniques:

(a) transforming recurrent networks in feedforward ones [2];
(b) introducing feedback connections [3];
(b) adding buffered context at the input [4];
(c) adding buffered context at the input and at the hidden layers [5].

Especially the last approach has given good results in speech recognition experiments [6]. All these new models are inherently limited in their representation of the past to a fixed period of time.

In this paper, a simple DMLNs (Dynamic Multi-Layered Network) is considered in which supervision is executed without considering a static input [7]. Instead of waiting for a fixed point, a learning algorithm is used in which the output supervision is done during the evolution of the activations. The learning environment is defined by a sequence of frames representing the natural time evolution of speech signals. The dynamic model considered is discrete instead of continuous and its transitions occur when a new frame is applied at the input.

The class of DMLNs utilized in this experiment is a simple one in which dynamic neurons, with feedback connections to themselves, have only incoming connections from the input layer.

METHOD

Instead of using classical "short-term" analysis approaches, like FFT, LPC or CEPSTRUM based filter bank, a physiologically-based joint synchrony/mean-rate Auditory Model (AM) of speech processing, proposed by S. Seneff [8], is considered as Digital Signal Processing front-end. Advantages of using an Auditory Model (**AM**) for speech recognition have been demonstrated in many contexts [9],[10].

A vector of 40+40 spectral-like parameters, representing the "mean rate" and the "synchronous" response of auditory neurons [8], is presented to the network each 2 ms, both during the learning and the testing phase. A block-diagram of the proposed AM is shown in Fig 1, while an example of the application of the AM DSP to the English sentence "Susan can't" (last consonants are omitted) is given in Fig. 2, in both "clean" and "noisy" speech condition [11]. The effectiveness of using this model is quite evident in Fig 2, observing the low frequency components of the two AM sonogram-like plots illustrated in (a) and (b), relatively to both "clean" (left) and "noisy" (right) speech.

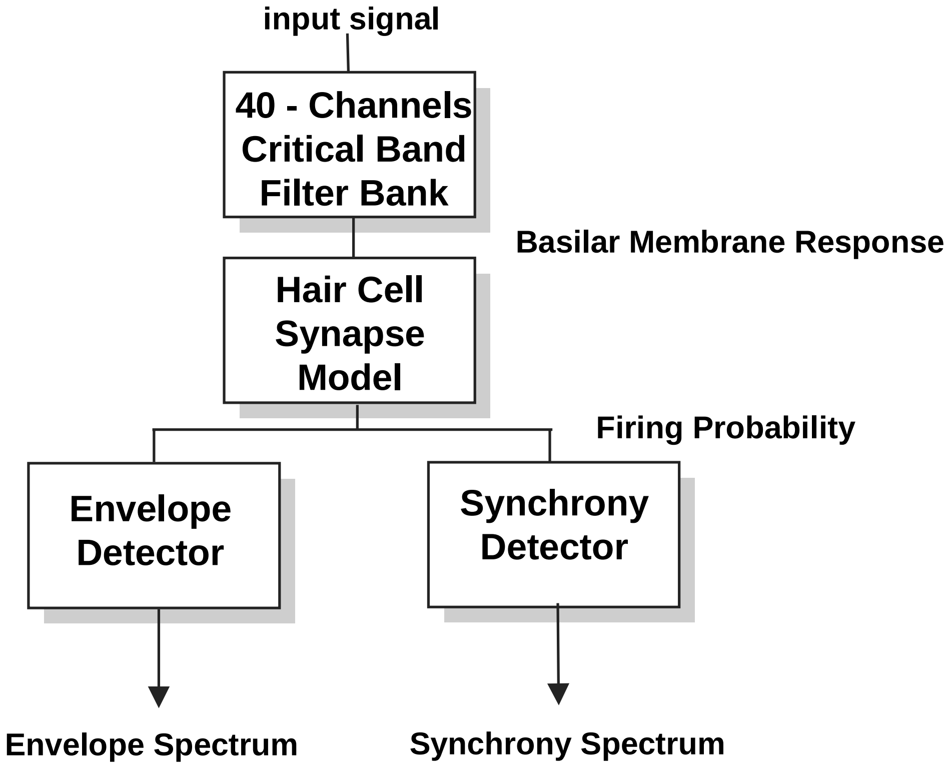


Fig.1. Block diagram of the joint synchrony/mean-rate auditory speech processing scheme.

Input stimuli are semi-automatically segmented by the use of a Multi-Level Segmentation (MLS) interactive algorithm [12] working on auditory model parameters. Advantages of using auditory models vs classical "short-term" analysis approaches for automatic speech segmentation have already been shown in litterature, especially in adverse conditions [11]. Various segmentation hypotheses are given by the MLS algorithm in form of a tree called "dendrogram" upon which the final segmentation is, up to now, manually extracted.. A graphic example of the output produced by the application of this algorithm to the same English sentence considered in Fig. 2 is illustrated in Fig. 3. As illustrated in Fig 3b, the same algorithm, applied to a "FFT-based spectrogram" instead of an "AM-based" one (Fig 3a), produces a more confusable segmentation "dendrogram", from which the final target segmentation is much more difficult to extract.

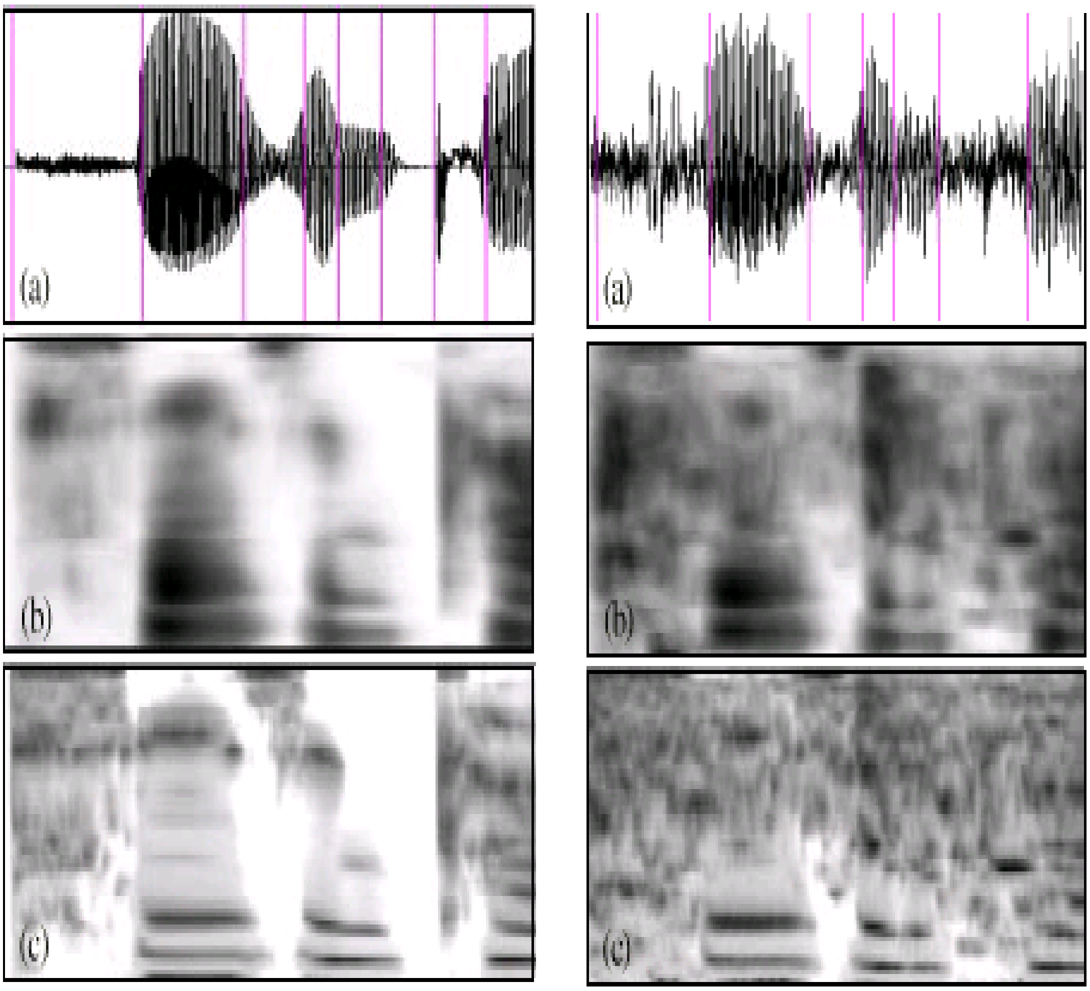


Fig. 2. "Auditory spectrogram". Envelope (b) and synchrony (c) outputs refer to the application of the joint synchrony/mean-rate auditory front-end to the English sentence "Susan can't" (last two consonants are omitted) spoken by a female speaker, both in "clean" (left) and "noisy" (right) condition.

As for supervision, during the learning phase, not only the knowledge of the stimulus identity is available, but also its fine segmentation characteristics.

Learning is organized within an isolated word recognition framework, and the network should output the right answer after presenting it each unknown stimulus. A dynamic recurrent neural network with local feedback connections, trained by EBPS [7], was used to discriminate input speech stimuli. In Fig 4 the structure of a generic Recurrent Network is shown, while in Fig 5 the mathematical framework of the recognition and the learning phase within the Extended Back Propagation for Sequences (EBPS) [7] theory is summarized.

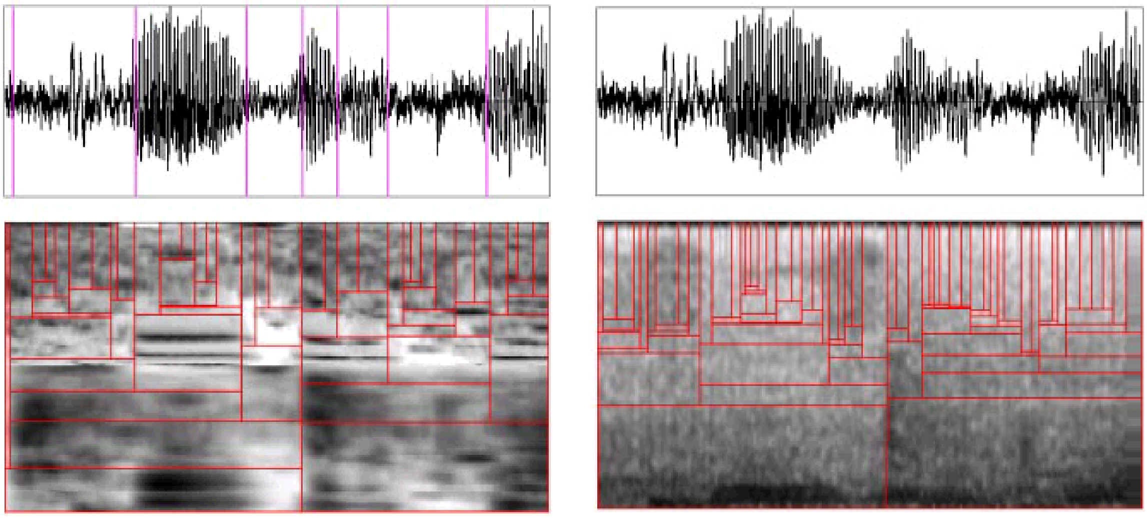
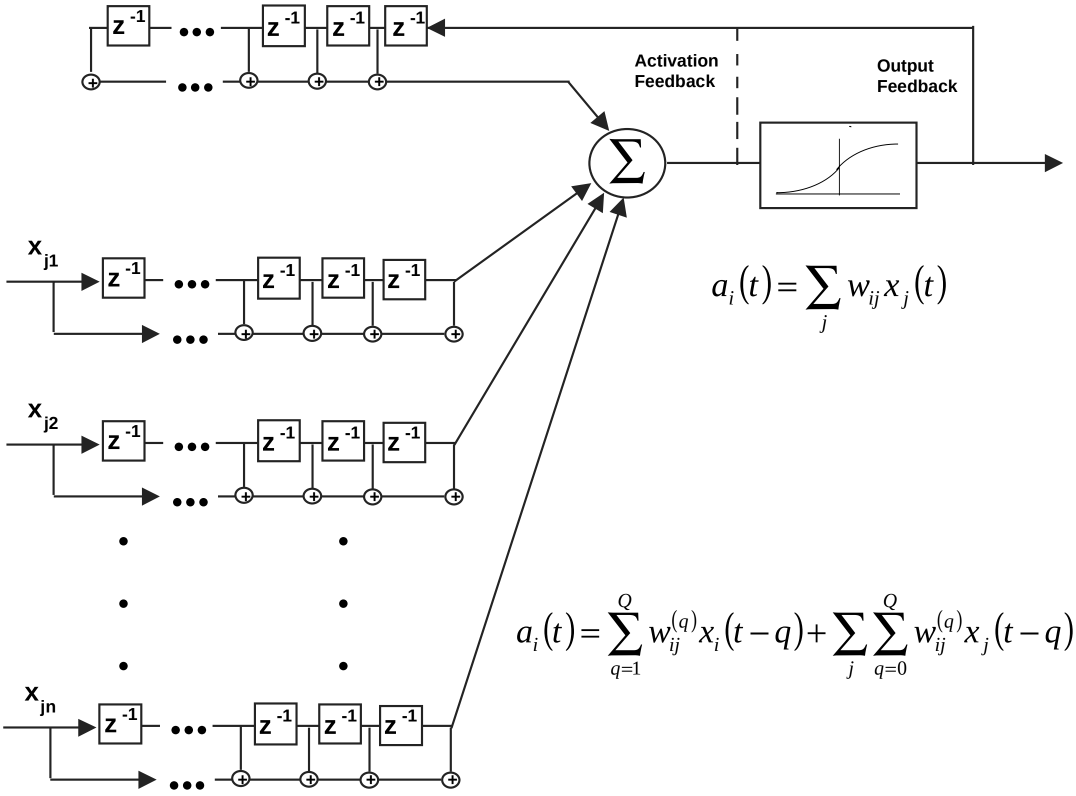


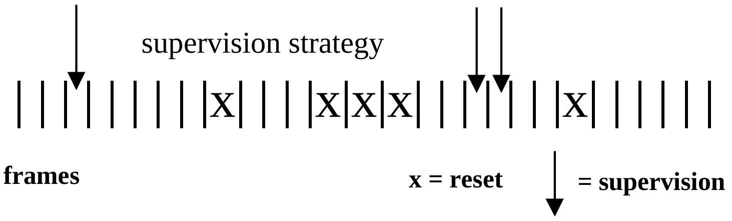
Fig. 3. Comparison of "dendrograms" produced by the àpplication of the MLS algorithm to the English noisy sentence "Susan ca(n't)" (last two consonants are omitted), using AM parameters (a) and FFT parameters (b).



Activation Feedback
Output Feedback
z^{-1}
x_{j1}
x_{j2}
x_{jn}
Σ
$a_i(t) = \sum_j w_{ij} x_j(t)$
$a_i(t) = \sum_{q=1}^{Q} w_{ij}^{(q)} x_i(t-q) + \sum_j \sum_{q=0}^{Q} w_{ij}^{(q)} x_j(t-q)$

Fig. 4. Structure of a generic Recurrent Neural Network (RNN).

algoritmo EBPS



$$\mathrm{C} = \sum_{t \in \Phi} \sigma_t C(t) = \sum_{t \in \Phi} \sum_{j \in O} \sigma_{tj} (x_j(t) - d_j(t))^2$$

σ_t, σ_{tj} *= supervision points*

Φ *=frame indexes*

O *= output nodes*

$$a_j(t) = \sum_{q=1}^{Q} w_{jj,q} x_j(t-q) + \sum_{i \in \Phi} \sum_{q=0}^{Q} w_{ji,q} x_i(t-q)$$

C-gradient

computed in a generic supervision time

$$-\frac{\partial C(\bar{t})}{\partial w_{ji,q}} = \frac{\partial C(\bar{t})}{\partial a_j(\bar{t})} \frac{\partial a_j(\bar{t})}{\partial w_{ji,q}} = \delta_i(\bar{t}) z_{ji,q}(\bar{t})$$

static neurons $z_{ji,0}(t) = x_i(t)$

dinamic neurons $z_{ji,r}(\bar{t}) = \sum_{q=1}^{Q} w_{ji,q} f'(a_j(\bar{t}-q)) z_{ji,r}(\bar{t}-q) + x_i(\bar{t}-r)$

Fig. 5. Mathematical framework of EBPS (Extended Back Propagation for Sequence) learning algorithm.

EXPERIMENTAL SETTING

The experiment described in this paper regards the automatic speaker independent recognition of the so called Italian I-set: /bi/, /tSi/, /di/, /dZi/, /i/, /pi/, /ti/, /vi/ plus other two i-like stimuli /Li/, /si/ (see SAMPA transcription [13]).

Speech signal is sampled at 16kHz, in a quiet office room, analyzed with a joint synchrony/mean-rate auditory-based signal processing [8] and successively segmented using a Multi-Level segmentation algorithm [12]. Fig 6 illustrated the computational environment in which all these steps are developed and implemented .

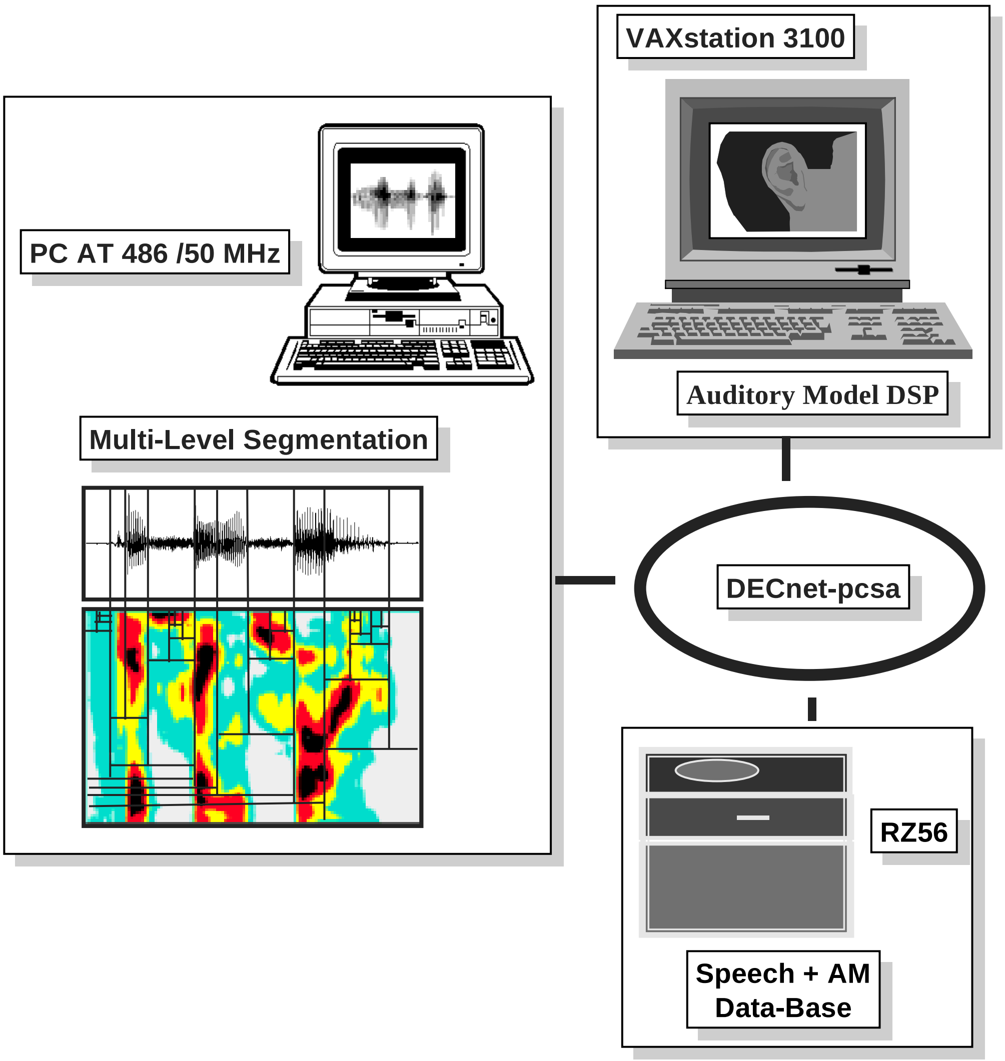


Fig. 6. Experimental Environment.

Speech data-base is made up of 7 male speakers characterized by 5 repetitions, for a total of 350 stimuli. Circularly one speaker is tested using the remaining 6 for learning.

The dynamic network utilized in this experiment has a MLN architecture in which both static and dynamic neurons cooperate. In particular a very simple DMLN structure is used, in which dynamic neurons, with a 4-delay feedbacks to themselves, have only incoming connections from the input layer. DMLN architecture is illustrated in Fig. 7 where also the structure of a dynamic neuron is drawn. 80 static neurons build the input level. They receive, frame by frame the output of the auditory front-end. 20 dynamic neurons with a 4-delay dynamic constitute the hidden layer and 10 static neurons, one for each phonetic stimulus, are considered at the output level.

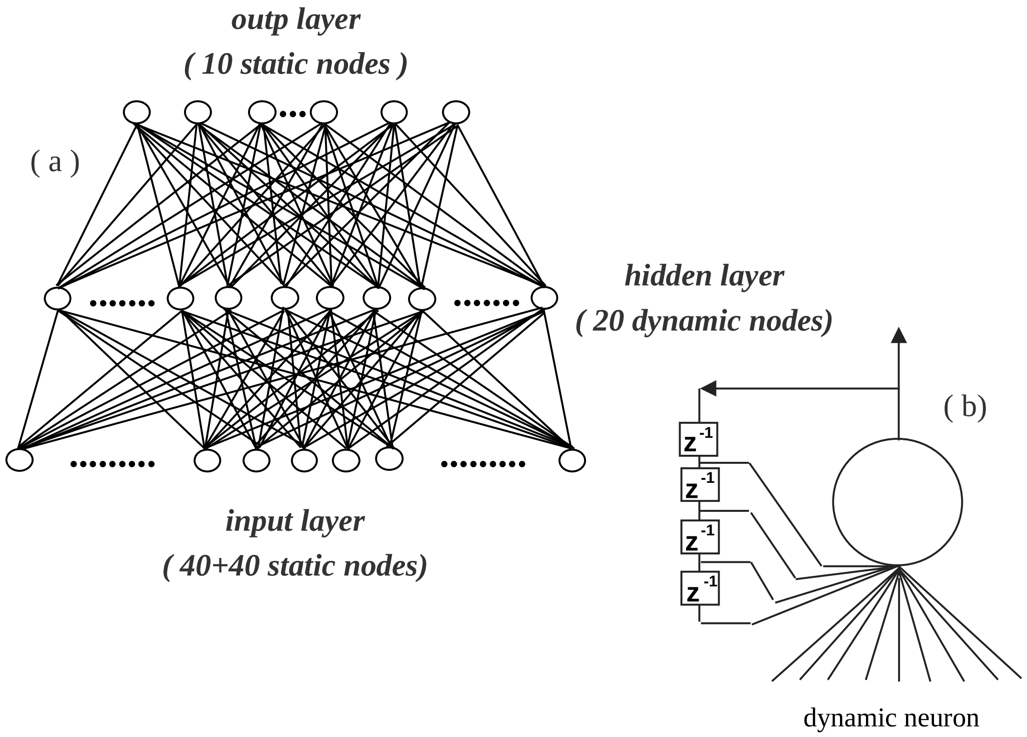


Fig. 7. Structure of the used Dynamic Multi-Layered Network and of a generic dynamic neuron.

Learning supervision is forced only at the penultimate frame of target phonetic stimuli as illustrated in Fig. 8 for the input stimulus /tSi/. This is obviously the worst situation in which all information given by formant transitions towards vowel /i/ is lost and only inplicit consonant characteristics are considered. In other words, considering plosives as an example, only burst information contributes to the final discrimination.

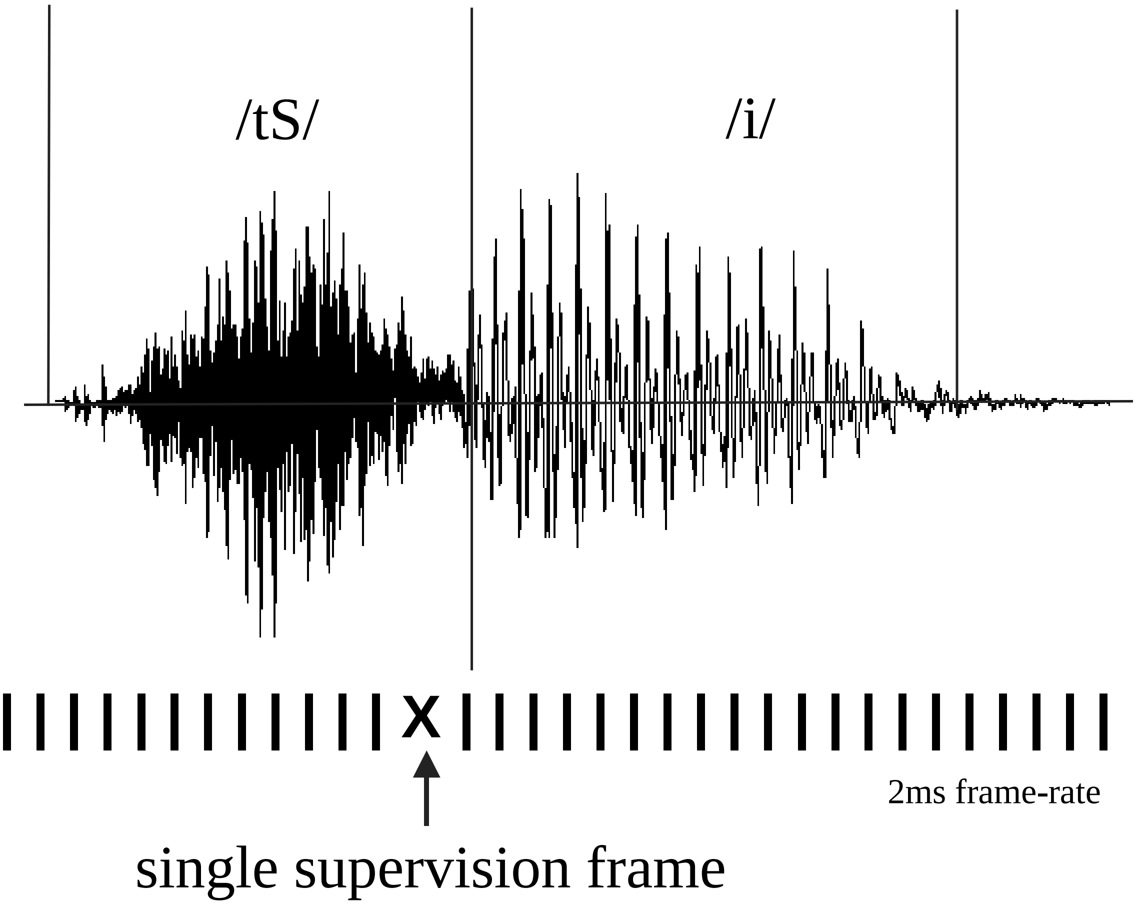


Fig. 8. Illustration of the learning supervision strategy. Only the penultimate frame of the target consonant is considered for supervision.

RESULTS

Given the low number of speakers results are given mediating 6 different experimental sets. Circularly one speaker is used as the test speaker, while the remaining 6 are used for training the DMLN. In Table I, all the different recognition error rates and the global mean rate are illustrated. Two speaker were particularly difficult to recognize, achieving only around 50% correct recognition rate, while all the others are rather good. The best speaker achieves 80% correct recognition rate. These results, given the effective difficulty of this task, are quite acceptable, but more experiments need to be exploited.

Speaker	Recognition Error
MM	22%
GF	32%
PT	36%
SR	48%
BC	48%
EP	36%
MR	27%
MEAN	35.6%

Tab. I. Recognition error rates for all speakers (remaining 6 speakers are used for learning).

FUTURE TRENDS

This work constitutes the base line, for other interesting experiments which are going to be developed. The Auditory Model DSP can obviously be enhanched and modified, while the Multi-Level segmentation algorithm can become completely automatic. Moreover the learning supervision strategy can be improved on behalf of more effective computational power. Another interesting field of research is represented by the following. Instead of learning by examples everything, a novel unified approach for integrating explicit knowledge and learning by examples in recurrent networks has been studied and successfully applied in ASR related problems [14]. In those experiments, characterized by isolated-word recognition tasks, lexical knowledge regarding the input words was injected into the recurrent network structure. The explicit knowledge is represented by automaton rules, which are directly injected into the connections of a network. This can be accomplished by using techniques based on linear programming instead of learning from random initial weights. Learning is conceived as a refinement process and is mainly responsible of uncertain information management. As well as lexical knowledge, explicit phonetic knowledge can be injected in the recurrent network architecture in order to facilitate the discrimination of phonetic stimuli. Moreover, with this structure, recognition can be accomplished not only in an isolated-word recognition framework but also in a continuous-speech recognition one.

REFERENCES

[1] D.E. Rumelhart, G.E. Hinton and R.J. Williams (1986), "Learning Internal Representations by ErrorPropagation" in *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, Vol. 1, Foundations, MIT Press, 1986, pp. 318-362.

[2] D.C. Plout and G.E. Hinton, (1987), "Learning Sets of Filters Using Backpropagation", *Computer Speech and Language,* Vol. 2 (2), July, pp. 35-61.

[3] F.J. Pineda, (1987), "Generalization of Back-Propagation to Recurrent Neural Networks", *Physical Review Letters*, Vol. 59, n. 19, Nov. 1987, pp. 2229-2232.

[4] T.J. Sejnowsky and C.R. Rosemberg, (1986), "NETTalk: a Parallel Network that Learns to Read Aloud", *Technical Report JHU/EECS-86/01*, 1986.

[5] A. Waibel, T. Hanazawa, G.E. Hinton, K. Shikano and K. Lang, (1987), "Phoneme Recognition Using Time-Delayed Neural Networks", *A.T.R. Technical Report TR-I-0006,* October 1987.

[6] A. Waibel, (1988), "Modularity in Neural Networks for Speech Recognition", *Proceedings of the 1988 IEEE Conference on Neural Information Processing Systems*, Denver CO, 1987.

[7] 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.

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

[9] J, R. Glass and V. W. Zue (1986), "Signal Representation for Acoustic Segmentation", *Proceedings First Australian Conference on Speech Science and Technology*, November 1986, pp. 124-129.

[10] M. J. Hunt and C. Lefebvre (1988), "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model, *Proceedings IEEE-ICASSP88*, New York, April 11-14, 1988, pp. 215-218.

[11] J. R. Glass and V. W. Zue (1988), "Multi-Level Acoustic Segmentation of Continuous Speech", *Proceedings of IEEE-ICASSP88*, New York, April 11-14, pp. 429-432.

[12] P. Cosi, "Ear Modelling for Speech Analysis and Recognition" (1992), *Proceedings of "Comparing Speech Signal Representations", ESCA Tutorial and Research Workshop*, Sheffield, England, 8-9 April 1992; paper ISSN 1018-4554 (to be published in J. Wiley & sons L.t.D. book).

[13] 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.

[14] P. Frasconi, M. Gori, M. Maggini and G. Soda (1991), "A Unified Approach for Integrating Explicit Knowledge and Learning by Example in Recurrent Networks", *Proceedings of IEEE-IJCNN91*, Seattle, 1991.