Experiments with Time Delay Networks and Dynamic Time Warping for speaker independent isolated digits recognition.

L. Bottou, F. Fogelman Soulié
EHEI, University of Paris V, PARIS

P. Blanchet, J.S. Liénard
LIMSI , ORSAY

ABSTRACT

We describe in this paper a speaker independent, global word recognition task using time delay networks. We first describe these networks as a way for learning feature extractors by constrained back-propagation. Such a time-delay network is shown to be capable of dealing with a test task: French digit recognition. The results are discussed and compared, on the same data sets, with those obtained with a classical time warping system. Both connectionist and classical systems achieved no more than 1% errors on the test set.

1- INTRODUCTION

Neural networks already have achieved encouraging results (Bridle, 1984) (Prager, 1986) (Kohonen, 1988) in a variety of tasks related to automatic speech recognition problems. Our aim here was to compare neural networks, and more precisely Time Delay Neural Networks (TDNN) to classical methods on a widely studied and well mastered task for today's speech recognition systems.

An efficient DTW system has been developed for some years at LIMSI (Gauvain, 1986), (Gauvain et al., 1983). Its performances have been shown to be state of the art on various data bases (Quenot et al., 1989). We thus compared our TDNNs and this DTW system on the same speaker independent digit recognition problem.

We propose in this paper a typical experiment of the capabilities of Time Delay Networks (TDNNs) with respect to DTW methods. We first describe the speech database we used. The time delay architecture is then depicted in the third section. In the fourth, we describe the experimental framework and comment the results achieved by the network. Comparison with the reference DTW experiments is provided in the fifth section.

2- DIGITS DATABASE

A speech data base, in French, has been elaborated at LIMSI. In the experiment reported here, we have only used part of the data base, namely the utterances of the 10 digits by 26 speakers, male (40%) and female. Each of the speaker pronounced each digit once.

We defined two different sets for the experiments:

- the learning set includes 16 speakers, with males and females in the same proportion as in the total set. We thus have 160 examples for learning.
- the test set includes the remaining 10 speakers (thus different from the 16 speakers used for training), which makes 100 examples for testing.

All the experiments were always based on the same learning and test sets: speakers were assigned to the two sets using alphabetical order, thus independently of any phonetical clue.

The signal has been processed in the following way, classically used at LIMSI (Gauvain, 1986), (Singer, 1988): the speech signal from the microphone has been filtered at 5 KHz through a low-pass filter, then sampled at 10 KHz with a 12 bits A/D converter. High frequency amplitudes are increased at 6 dB per octave. A DFT is applied on successive 25.6 ms time frames, overlapping by 12.8 ms. Thus 128 energy spectra values are generated in the 0-5 K Hz frequency domain. A Bark scaled 16-channels filterbank is then simulated by averaging on triangular frequency windows. The energy spectra are then log-compressed.

This processing thus results in coding the speech signal into sixteen eight bits values per 12.8 ms time frame.

3- TIME-DELAY NETWORKS

Our preferred way to describe time-delay networks consists to show how the Gradient Back Propagation (GBP) rule may be used in multi-layer perceptrons (MLPs) for discovering time-invariant feature extractors.

Let us describe now a network for learning those feature extractors. Each feature extractor is built from a set of hidden units. These units are locally connected to a window scanning the input data. The simplest way to give a time invariant behavior to these units is to insure that all their incoming weights will remain identical during the training phase. Such extensions of the standard back-propagation algorithm were discussed in the PDP book (Rumelhart, 1986). The general theme of constrained back-propagation has also been extensively studied in (Le Cun, 1988).

Moreover, the time invariant features may be used as input data for another layer of feature extractors and so on. The entire network thus can be trained by constrained back-propagation. As a side effect, some connections are drawn between units corresponding to different times. Such networks are thus called Time Delay Networks.

Experiments have been run (Lang, 1988) to compare fully connected, locally connected and TDNN networks. The experiments have been carried out on the /b/,/d/,/e/,/v/ task. The results show, even on this very simple task, that the time-delay trick is quite appropriate. Other experiments at ATR (Waibel, 1987) showed that such networks were capable of achieving better results than a Hidden Markov Model (HMM) on a japanese /b/,/d/,/g/ recognition task.

How do architectures, learning time, and performances scale with the complexity of the task ? In our digit data base, the speech signal lasts about one second, which is the input to the network. This is to be compared with e.g. the /b/,/d/,/g/ problem (Waibel, 1987) where the typical speech data lasted 150 ms only: with our larger framing rate, this means a factor of 4 in the number of input units. We attempt here to reduce the number of cells in the hidden layers by progressively reducing the number of the cells in the feature extractors. However, our network is about 1300 cells large, where Waibel's was about 400.

We have one 16 dimensional vector as input every time slice (fig 1). A first layer of 8 feature extractors operating on windows of three consecutive vectors transforms these inputs into one 8 dimensional vector every two time-slices. A new layer of 8 feature extractors, windowed on seven consecutive vectors, give one 8 dimensional vector every ten time slices. The resulting vectors are then fully connected to 10 decision cells, one for each digit to be recognized.

4- RESULTS WITH THE NETWORK

For training the network, we performed some additional processing. The input layer was set with 65 time frames. We built 640 patterns out of 160 training utterances: each utterance is randomly shifted in the 128 first ms of the 832ms window, simulating a poor word segmentation. This is repeated 4 times, which leads to 4 patterns per utterance. The spectrogram energies were linearly scaled into the [-1,+1] interval, independently for each speaker.

In the same way, with only 2 random shifts, we have built 200 patterns out of the 100 test utterances. Of course, the test speakers are not the same than the training speakers.

The network was trained using an all purpose back-propagation algorithm, exactly as described in (Fogelman, 1987) (Le Cun, 1987).

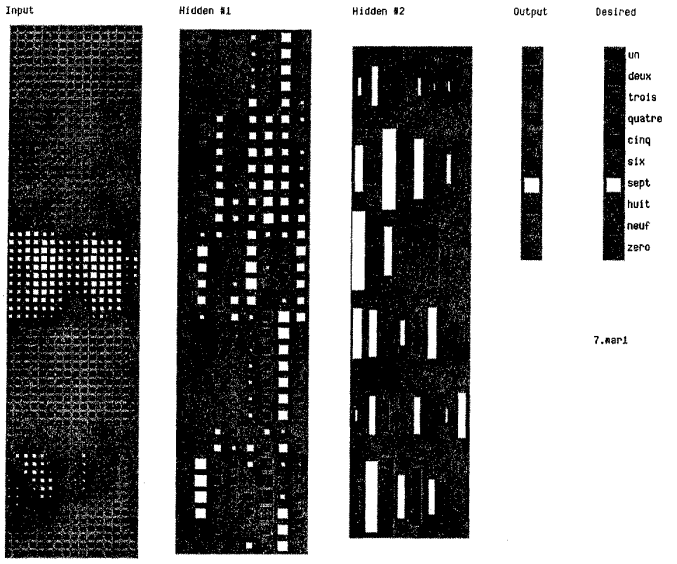

figure 1- Diagram of the Time Delay Network.

We ran the learning task about thirty times. We always stopped the simulation after 30 sweeps of the 640 training patterns (i.e. 90 minutes on a Sun4 workstation). The network never achieved less than 98% correct answers on the training set and 94% on the test set.

The best run produced a network able to correctly classify 99.21% of the training patterns and 99% of the test patterns (i.e. 1 unrecognized word out of the 100 test utterances). Unfortunately, our speech data base is clearly too small for really validating such a performance. However, we reproduced a couple of times this result with different initial weights. It is interesting to notice that after 6 sweeps, the network already achieved 97.9% on the training set and 93% on the test set, and after 15 sweeps, 98.3% and 98%. (see fig.2)

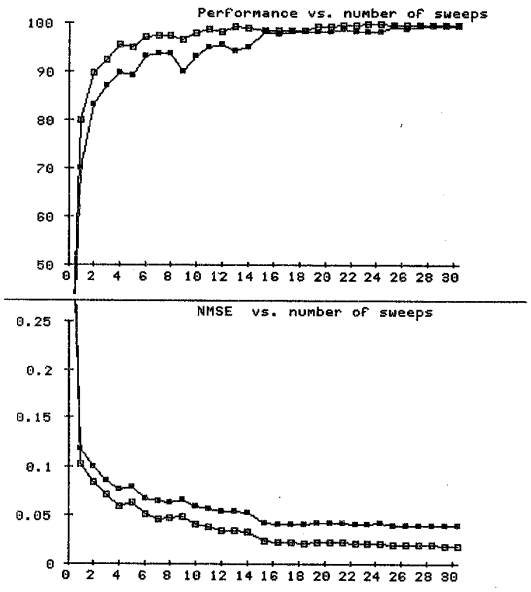

figure 2- Error and Performances curves, for both the training set and the test set.

These results show that a small number of epochs is sufficient for efficiently training a time delay network, without problem specific or machine dependent tricks.

However, the error made by the network is clearly not acceptable. A 9 ("neuf") had been recognized as a 4 ("quatre"), which is rather disappointing at first sight.

We attempted to visually identify significant information in the first layer weights. For example, it seems easy to recognize formant's movements detectors or other phonetic features detectors. But we are unable to really distinguish between the randomly produced ones and those that the network effectively uses for classifying the patterns.

We thus preferred to study the activities of the last hidden layer cells: there are 6x8 such cells, whose activities can be viewed as 6 vectors in an 8-dimension space (see fig 1). Each of these vectors represents the activity of the 8 masks in 6 different portions of the signal: the set of the activity vectors for all the available patterns (training and test sets) is vector quantized through a k-means clustering technique. The result is a description of each digit as a sequence of sub-word units. In figure 3, for example, we show the results of a clustering of some patterns in 8 clusters: the network has clearly extracted a rather stable decomposition of each digit in sub-word units, of a slightly larger grain than phonemes. (cluster 8 appears to be the prototype code for "silence")

UN	1	1	4	8	8	8		SIX	6	6	6	6	6	8
UN	3	1	4	8	8	8		SIX	6	5	6	8	8	8
UN	1	1	4	8	8	8		SIX	6	6	6	6	6	8
UN	1	1	4	8	8	8		SIX	6	6	6	6	6	6
DEUX	7	2	2	8	8	8		SEPT	6	7	7	8	8	8
DEUX	7	2	2	8	8	8		SEPT	6	7	2	8	8	8
DEUX	7	2	2	8	8	8		SEPT	6	7	2	8	6	8
DEUX	7	2	2	8	8	8		SEPT	6	7	2	8	8	8
TROIS	7	3	4	4	8	8		HUIT	5	6	4	8	8	8
TROIS	8	3	3	4	8	8		HUIT	5	6	4	8	8	8
TROIS	3	3	4	8	8	8		HUIT	5	6	4	8	8	8
TROIS	7	3	3	4	8	8		HUIT	5	5	6	8	6	8
QUATRE	2	1	4	4	8	8		NEUF	7	2	1	8	8	8
QUATRE	7	4	4	4	8	8		NEUF	2	1	6	8	8	8
QUATRE	7	1	4	4	8	8		NEUF	7	2	2	8	8	8
QUATRE	7	1	4	4	8	8		NEUF	2	1	6	8	8	8
CINQ	6	3	2	8	6	8		ZERO	5	2	3	8	8	8
CINQ	6	7	1	7	6	8		ZERO	5	2	3	3	8	8
CINQ	6	8	1	2	6	8		ZERO	6	5	3	8	8	8
CINQ	6	7	1	1	8	8		ZERO	7	2	3	3	3	8

figure 3 - Digits decomposition extracted by the network.

Moreover, this clustering technique also shows that the decomposition of the 9 that the network has extracted is the least stable. This remark could give some insight to start understanding the origin of the 9 versus 4 error.

Of course, we do not claim that the results obtained so-far actually achieve the extraction of sub-word units which are completely understandable and meaningful. However, we think that following up this line of investigation could provide significant results in that area.

5- DYNAMIC TIME WARPING

In order to compare the previous results with classical systems ones, we used the LIMSI's DTW system described in (Gauvain, 1983) (Mariani, 1983) and (Singer, 1988) on the same training and test utterances, processed as follows.

Each time frame of the spectrogram log-compressed energies is first normalized. A cosine transformation is then applied for extracting 8 cepstral coefficients. The time warping process is performed using these 8 parameters and the mean energy. The sixteen learning utterances for each word are precisely segmented , then averaged along their time warping path, using an algorithm described in (Singer, 1988). The resulting references require approximatively as much memory space as the TDNNs weights.

The 100 test utterances have been presented and 99 were correctly classified. The error source was identified as a bad segmented training reference. We did not try to check the DTW performance using data as badly segmented as those used for the TDNN experiment. The DTW system seems indeed very sensitive to training data segmentation, however, real training data may quite well segmented in practice.

We reach here the limits of objectivity while comparing two fundamentally different techniques on the same task. Neither very precisely segmented, nor strongly unsegmented data are better fitted to the reality. Experimental constraints may insure or forbid such a precise segmentation.

6- CONCLUSION

We have presented here a Time Delay Neural Network trained on a small French digit recognition task. Our work clearly demonstrates that adequate neural nets have many significant abilities for speaker independent speech recognition tasks. Our best network achieved about the same level of results than a well tuned classical system. We think that this is rather encouraging for the Neural Net approach, which has not been yet fully explored nor optimized. In addition, the TDNN has shown its ability to learn on badly segmented references, which suggests that a phonetical speech recognition system could be trained with affordable coarsely segmented databases.

We tried to compare TDNNs and DTW as far as possible. But fundamentally different techniques are best measured by fundamentally different tests. Furthermore, cepstral coefficients are known as a good parameter set for improving speaker independence in a time warping process. Very little is known about the respective properties of different signal processing techniques for feeding a neural net. Spectrograms appeared as the easier way for evaluating networks, but it might be possible that further research will enable to improve MLPs 'performances by using more adequate processing.

Finally, we have given some hints to utilize Time Delay Networks for sub-words units extraction. More investigation is needed to get significant results. However, we think that the results obtained so far are an incentive to go further in the direction presented here.

7- ACKNOWLEDGEMENTS

This work was partly supported by Brain project n°ST2J-0418-C. L.B. is supported by DRET grant n° 87/808/19 and P.B. by DRET grant n°87/100.

One of us (L.B.) gratefully acknowledges helpful conversations with G. Hinton and Y. Le Cun, during a stay at the University of Toronto.

8- REFERENCES

L. BOTTOU: Reconnaissance de la parole par réseaux multi-couches. In "Neuro-Nîmes 88" pp371-382, EC2 eds (1988).

L.BOTTOU, F.FOGELMAN, P.BLANCHET, J.S LIENARD: Speaker Independent isolated word recognition: Multi-Layer Perceptrons vs Dynamic Time Warping, (1989) to be published in Neural Networks.

H. BOURLARD, C.J. WELLEKENS: Multilayer perceptrons and automatic speech recognition. In IEEE First International Conference on Neural Networks, San Diego, june 1987, IEEE catalog n° 87TH0191-7, vol. IV-pp407-416, (1987).

J.S. BRIDLE, R.K. MOORE: Boltzmann Machines for speech pattern processing. Proc. Inst. Acoust. Autumn Meeting, (1984).

J.L. ELMAN, D. ZIPSER: Learning the hidden structure of speech. Tech. Report, Univ. of California, San Diego, (feb. 1987).

F. FOGELMAN SOULIE, P. GALLINARI, Y. LE CUN, S. THIRIA: Evaluation of network architectures on test learning tasks. IEEE First International Conference on Neural Networks, San Diego, june 1987, IEEE catalog n° 87TH0191-7, vol. II pp653-660, (1987).

F. FOGELMAN SOULIE, P. GALLINARI, Y. LE CUN, S. THIRIA: Network learning, In "Machine Learning ", vol 3, Y. Kodratoff, R. Michalski eds, Morgan Kauffman, (1989).

J.L. GAUVAIN: A syllable-based isolated word recognition experiment. In Proc. IEEE ICASSP-86, (1986).

J.L. GAUVAIN, J.MARIANI, J.S. LIENARD: On the use of time compression for word-based speech recognition - IEEE - ICASSP - Boston, (1983).

F. JELINEK, L.R. BAHL, R.L.MERCER Continuous Speech Recognition: statistical methods. Handbook of Statistics II, H.P. Krishnaiad Ed., North Holland, (1982).

T. KOHONEN: The "Neural" phonetic typewriter. IEEE Computer, 11-22, (March 1988).

K. LANG, G.E. HINTON: The development of TDNN architecture for speech recognition. Tech. Report CMU-CS-88-152, (1988).

Y. LE CUN: Modèles connexionnistes de l'apprentissage, Thèse, Paris, (1987).

Y. LE CUN: A theoretical framework for back-propagation, In "Connectionist Models: a summer school", D.Touretzky (ed), Morgan-Kaufmann. pp21-28 (1988).

S.E. LEVINSON: Structural methods in automatic speech recognition. Proc. of the IEEE, vol. 73, n°11, (1985).

J. MARIANI, B. PROUTS, J.L. GAUVAIN, J.T. GANGOLF : Man Machine Speech Communication Systems, including word-based recognition and text to speech synthesis. IFIP World Computer Congress, Paris, (1983).

R.W. PRAGER, T.D. HARRISON, F. FALLSIDE: Boltzmann machines for speech recognition. Computer, Speech and Language, voll N°1, 3-27, (1986).

G. QUENOT, J.L. GAUVAIN, J.J. GANGOLF, J.J. MARIANI: A dynamic programming processor for speech recognition. In IEEE Jl of Solid State Circuit, vol. 24, n°2, (1989).

D.E. RUMELHART, G.E. HINTON, R.J. WILLIAMS: Learning internal representations by error propagation. In "Parallel distributed processing", D.E. Rumelhart, J.L.McClelland eds, MIT Press, vol 1, 318-362, (1986).

H. SINGER: Utilisation de dissylabes pour la reconnaissance de la parole. Rapport LIMSI, 88-4, (1988).

A. WAIBEL, T. HANAZAWA, G. HINTON, K. SHIKANO, K. LANG: Phoneme recognition: neural networks vs. Hidden Markov Models. Proceedings ICASSP 88, S-Vol.1, 107-110, (1988).

A. WAIBEL, T. HANAZAWA, G. HINTON, K. SHIKANO, K. LANG: Phoneme recogntion using Time-Delay Neural Networks. Preprint ATR Interpreting Telephony research Laboratories, (oct 1987).