A Comparative Study of the Multi-Layer Perceptron, the Multi-Output Layer Perceptron, the Time-Delay Neural Network and the Kohonen Self-Organising Map in an Automatic Speech Recognition Task

F J Owens, R Andonie*, G H Zheng, A Cataron, S Manciulea****

School of Electrical & Mechanical Engineering, University of Ulster,
Jordanstown, Newtownabbey, Co Antrim, BT 37 0QB, Northern Ireland
** Computer Science Department, Wayne State University, Detroit, MI 48202, USA*
*** Dept. of Electronics and Computers, Transylvania University, 2200 Brasov, Romania.*

Abstract

This paper describes a study of the use of four different neural network techniques for automatic speech recognition (ASR) using two common, real-world application databases. The neural network techniques investigated were the Multi-Layer Perceptron (MLP), the Multi-Output-Layer Perceptron (MOLP), which is an improved version of the MLP , the Time-Delay Neural Network (TDNN) and the Kohohnen Self-Organising Map (SOM). The speech test data consisted of a clean database, acquired in a relatively noise-free room environment, and a telephone database, acquired over conventional dial-up lines. Each database comprised 20 repetitions of 12 isolated words (the digits 0 - 9 plus 'nought' and 'oh') each spoken by 25 talkers. Each word was parameterised into a time sequence of 15 frames of an 18-dimension feature vector, consisting of 8 Mel-frequency Cepstral Coefficients (MFCCs), the corresponding frame-to-frame MFCC differential coefficients and absolute and differential signal energy coefficients. In a speaker-independent, isolated-word speech recognition task, the respective recognition scores for the MLP, MOLP, TDNN and SOM were 93.2%, 95.5%, 95.1%, and 97.1% respectively for the clean speech database , and 76.3%, 90.5%, 90.5% and 96.8% respectively for the telephone database.

1. Introduction

The challenging computational problems associated with speech recognition and the limited success of the conventional pattern recognition techniques proposed to solve them have fostered the development of neural network approaches to speech recognition tasks. The proposals made in the literature differ mainly in how the speech signals are converted to a format which can be used as a neural network input, what the network should recognize (i.e., speaker-independent or speaker-dependent recognition of words or phonemes), and what type of neural network is used.

For a neural network to be dynamic and thus able to process speech information, it must be given memory [Elm90]. The most simple strategy is to represent a sequence of incoming data "simultaneously" on the input layer of the neural network. This is a *static* strategy since it does not explicitly address the temporal nature of the data. The input layer is a buffer which holds the current temporal data for processing. There are two basic ways to change the content of the buffer: *i)* the buffer acts like a shift register; *ii)* the buffer contains data sampled within the current time window (windows may overlap). The network typically used is a feedforward backpropagation network. This approach has been taken by Bengio *et al.* [Ben89] and Freisleben *et al.* [Frei93]. Owens *et al.* investigated the use of multi-layer perceptrons and "multi-output-layer perceptrons" for automatic speech recognition [Owe96]. Both of these approaches are also static. The static approach has several disavantages [Elm90]: it imposes a fixed duration for patterns, it does not distinguish between absolute and relative temporal positions, and the backpropagation network usually does not handle novel inputs well.

Time-delay neural networks are a group of neural networks in which the input signal is considered together with delayed versions of it (i.e., the output of the network depends on its current and previous inputs). A reference work describing the use of time-delay neural networks in phoneme recognition is that of Waibel *et al.* [Wai89]. It is notable that their model has a desirable property related to the

Proceedings of the EIS'98 International ICSC Symposium on Engineering of Intelligent Systems, ICSC Academic Press, Tenerife, Spain, February 11-13, 1998, 624-629.

dynamic structure of speech: it is translation invariant, that is, the features learned by the network are insensitive to shifts in time. Bottou *et al.* applied a time-delay neural network to the task of speaker-independent isolated digit recognition with very good results [Bott90].

It has been found that a feedforward network is unable to learn temporal relationship and it must be programmed in advance [Fu94]. On the other hand, recurrent neural networks hold great promise in speech recognition. They can store temporal information and somehow manage to learn temporal relationship. Generally speaking, recurrent networks can learn complex structures involving precedence (not necessarily temporal) relationship. Another attractive point is that recurrent networks use a reduced number of neurons compared to static and time-delay neural networks. This is due to the fact that static and time-delay networks use a spatial representation of temporal patterns while recurrent networks use temporal representation, which means that a temporal sequence enters the network one data element at a time. Several recurrent networks have been used for speech recognition. For instance, Hopfield neural networks were applied for vowel recognition [Gar93].

A very different neural network model used in automatic speech recognition is the self-organizing feature map. Kohonen [Koh88] employed the "phonotopic map", which is based on such a model, in a speech transcription system implemented in a PC environment. This system is used as a "phonetic typewriter" that can produce text from arbitrary dictation. Further extensions of the use of self-organizing feature maps in speech recognition can be found in [Tatt90] and [Beau93].

The most of these cited papers report good recognition rates compared to other approaches. For instance, Freisleben *et al.* presented a speech recognition system that allows recognition of a limited vocabulary of spoken words (45 German words) in a speaker-independent manner [Frei93]. Their experiments have shown that the recognition rate is up to 91% for unknown speakers of the same sex and up to 72% for a mix of both male and female speakers. It is difficult to compare directly the results obtained by different authors since these results usually concern particular speech input data. Therefore, it is desirable to consider the same input data for different neural network approaches to speech recognition. This could give us a greater insight about the performance of these approaches.

The aim of the paper is to compare four different neural network techniques in speech recognition. Sections 2 and 3 describe briefly the speech databases and preprocessing we use. In Sections 4-6 we present the architecture and training of the neural models we use: two static models (the multi-layer perceptron and the multi-output-layer perceptron), a time-delay neural network, and a self-organizing feature map. The numerical results of the simulations, are concentrated in Section 7. Some conclusions are presented in Section 8.

2. Speech Databases

To test different artificial neural network learning algorithms, two speech databases were used in the speech and speaker recognition experiments and each consisted of 20 repetitions of each of 12 words by each of 25 talkers, giving a total database size of 6000 utterances. The first database was provided by 19 males and 6 females. The second database was provided by 15 males and 10 females. In each case the age range of the talkers was from 20 to 60 years and all had a Northern Irish accent. The words chosen were the spoken digits, 'one' to 'nine', plus 'nought', 'oh' and 'zero'. The first database was intended to be a 'clean' speech database, with data collected under controlled conditions and a minimum amount of noise interference etc. Each talker's speech was recorded, in a single session, on to conventional audio-cassettes using a high-quality microphone and a professional cassette tape recorder. The tapes were then mounted in a cassette deck and the speech signal was pre-amplified before being passed through a 4th order, 3.5kHz Butterworth anti-aliasing filter. The signal was then amplified and digitised using a 12-bit, analogue-to-digital converter (A/D) operating at a 7.5kHz sampling rate. The digitised speech was then manually end-pointed and stored on disk.

The second database was intended to more closely represent the expected conditions of operation in a 'real-world' application of speech/speaker recognition using the telephone network. For the telephone database, the speech was acquired and stored in *real-time*. It was possible to enter speech data from any location which was equipped with a conventional telephone and a terminal which could access the University of Ulster computer network. The terminal was used to 'log-on' to the acquisition computer and a connection was also established to the computer via the local dialled-up telephone lines and the telephone interface. The telephone handset type used in every case was the BT 'Tribune', and each talker's speech data was

recorded in a single session. The speaker was then prompted via the terminal on the computer with the words to be spoken. The telephone signal was interfaced to the same signal pre-processing and A/D circuitry as used for the 'clean' speech database. However, in this case, the digitised speech was automatically end-pointed by the computer using simple thresholding of speech signal energy and zero-crossing rate measurements.

3. Front-End Processing

Speech pattern extraction was based on a fast Fourier transform (FFT) - based mel-scale (non-linear frequency scale) filterbank. The spectral output from the filterbank was transformed to the cepstral domain using a discrete cosine transform (DCT). The input speech was split into 20ms frames using an overlapped Hamming window of duration 30ms and a standard radix-2 decimation-in-time FFT algorithm was used for computing the short-time spectrum. The mel-scale filterbank outputs Xj were computed by multiplying the short-time magnitude spectrum using the equi-spaced, triangular mel-scale filterbank and aggregating the weighted spectral components falling within each band. The mel frequency scale is related to the normal frequency scale using the relation

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \qquad (1)$$

The MFCC coefficients C_i were computed from the log filterbank outputs Xj using the following DCT relation

$$C_i = \sum_{j=1}^{N} X_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right), \quad 1 \le i \le M \qquad (2)$$

where N is the number of filterbank channels and M is the number of desired cepstral coefficients. Each delta coefficient, d_P , for frame P were computed using the expression

$$d_P = \frac{\sum_{i=1}^{M} i(C_{P+i} - C_{P-i})}{2\sum_{i=1}^{M} i^2} \qquad (3)$$

where C_P is the cepstral coefficient for frame P. At the beginning and end of the utterance, simple first-order differences were used, that is

$$d_P = C_{P+1} - C_P , \quad P < M$$

and (4)

$$d_P = C_P - C_{P-1} , \quad P \ge N_F - M$$

where N_F is the number of frame vectors in the word.

The values of N and M used were 16 and 8 respectively, giving a frame pattern vector of dimension 18, consisting of 8 absolute MFCC coefficients, 1 log-energy coefficient and their 9 related delta coefficients.

4. The Multi-Output-Layer Perceptron (MOLP)

The multi-output-layer perceptron (MOLP) is a relatively new type of network defined by Zheng and Owens [Zhe93]. The MOLP is an MLP with multiple layers of output nodes as shown in Fig. 1. Each additional output layer owns the same nodes as the first output layer. Every node in each additional layer is directly connected to every node in the last hidden layer and corresponding node in the previous output layer(s). The network may be trained in both constructive and non-constructive ways [Zhe96]. In non-constructive training the number of output layers is pre-determined and the errors from each output layer are back-propagated simultaneously to the hidden layer. In constructive learning, the starting point is a conventional multi-layer feedforward architecture to which additional output layers are progressively added. In constructive serial learning the original network weights are kept frozen and only the weights in the new output layer are trained. In constructive parallel learning, all of the previously added output layer weights are further trained while the new output layer weights are being trained. It has been found that the best performance in an automatic

speech recognition application is obtained using a non-constructive learning procedure [Zhe96].

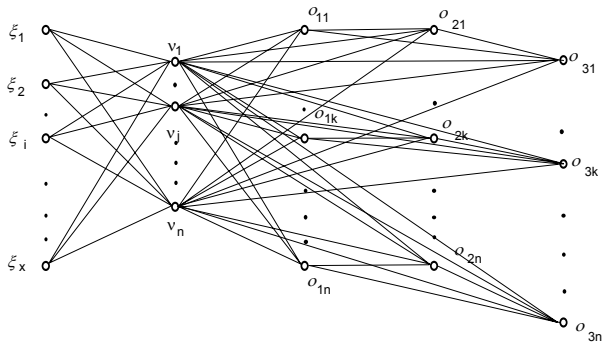

Fig. 1 : A Three-Output-Layer and One-Hidden-Layer MOLP Network

The MLP and MOLP networks had 270 (18 × 15) input nodes, and there were 12 output nodes, one node for each word. Output node target values of 0.1 (logic value 0) and 0.9 (logic value 1) were used. During training, the threshold values used for assessing correct classification were 0.2 (0) and 0.8 (1), and during recognition the corresponding values were 0.3 and 0.7, respectively. The backpropagation algorithm was used and task-optimized values of 0.2 and 0.8 were used for the learning rate and momentum terms, respectively.

Previous experimental investigations showed that the classification ability of each MOLP is superior to that of an equivalent conventional MLP network [Owe96]. In general, this performance increase can be achived with shorter training times and simpler network architectures.

In our comparative study we used the MOLP model with non-constructive learning and the conventional MLP as reference models of the static strategy.

5. The Time-Delay Neural Model

Time-delay neural networks (TDNNs) are particular feedforward networks having a memory of the input patterns [Wai89]. For the speech recognition approach, it is usual to have large input patterns. TDNNs allow the input layer to be smaller than the input pattern dimension. A window is moved over the input pattern. The window is as large as the number of neurons on the input layer. Each element of the input pattern must belong to at least one window. The data in a window is the input data for the neural network. The backpropagation algorithm used in TDNN learning consists of a few steps. First of all, as many copies of the network are made as the number of windows. A copy of the network uses the data in a window. Then, for each copy, a classical backpropagation step is performed. The average adjustements of the corresponding weights are the adjustments of the weights of the original network.

Our experiments showed that the networks with one hidden layer gave better results than using two or more hidden layers. Due to the database organisation, every neural network has 18x15 neurons on the input layer (i.e., 15 frames, every frame having 18 coefficients). The best results in word recognition were obtained with a network with 15 feature units and the total delay length equal to 7 on the hidden layer (15x7).

6. Self-Organizing Maps

The self-organizing map (SOM) represents the result of a quantization algorithm that places a number of codebook vectors into a high-dimensional input data space to approximate to its data sets in an ordered fashion [Koh89]. When local-order relations are defined between the reference vectors, the relative values of the latter are made to depend on each other as if their neighbouring values would lie along an “elastic surface”. By means of the self-organizing algorithm, this “surface” becomes defined as a kind of nonlinear regression of the reference vectors through the data points. A mapping from a high-dimensional data space $\Re^n$ onto, say, a two-dimensional lattice of points is thereby also defined. Such a mapping can effectively be used to visualize metric ordering relations of input samples. In practice, the mapping is obtained as an asymptotic state in an unsupervised learning process. Like any unsupervised classification method, it may also be used to find clusters in the input data and to identify an unknown data vector with one of the clusters.

We used one map for each class, an idea inspired and adapted from [Vee95]. Each map is a hexagonal lattice with 17 × 15 nodes, this size being determined experimentally. A 270-dimensional vector is too large for this neural model, considering both processing time and recognition performance. Therefore, we had to split up all the data vectors, considering the 15 initial frames. This way we reduced the dimensionality of the input vector from 270 to 18 (each frame contains 18

coefficients). We performed the following steps for creating the training and test data files:

For training purposes, from the 15 frames of each word, 5 frames were picked up at random. This means that for training, 30% of the available samples were chosen. This was done for each class. Each map was trained using about 10,000 steps in the first stage (the ordering phase) and about 120,000 steps in the second phase. In order to verify the recognition capability of the system, we used all patterns from each class. The input frames were introduced into the network in an ordered fashion: 15 frames were presented sequentially to the each network. A frame was deemed to belong to a map if and only if it produced the smallest quantization error. If the majority of the 15 frames were assigned to a neural network *k*, then the example was deemed to belong to class *k*. It turned out that, for particular cases, more than one map had the same number of assigned frames. Therefore, this model can't always avoid ambiguities in classification. For such ambiguous cases, the overall quantization error is computed (i.e., for each network the quantization errors are summed) and these errors are compared for taking a decision.

7. Summary of Experimental Results

MLP, MOLP, TDNN and SOM summary generalisation characteristics for automatic speaker independent word recognition using the 'clean' speech database are shown in Table 1.

MLP	*MOLP*	*TDNN*	*SOM*
93.3%	95.5%	95.1%	97.1%

Table 1 : ASR Results using the 'Clean' Speech Database

MOLP, TDNN, and SOM summary generalisation characteristics for automatic speaker independent word recognition using the telephone speech database are shown in Table 2.

MLP	*MOLP*	*TDNN*	*SOM*
76.3%	90.5	90.5%	96.8%

Table 2 : ASR Results using the Telephone Speech Database

8. Conclusions

The performance of the four considered neural techniques appears to be quite competitive to other results reported in the literature, e.g., Hidden Markov Modelling (HMM) and Dynamic Time Warping (DTW). The SOM neural model proved to be the most accurate speech classifier. Considering the processing time for training, the SOM model is faster than the TDNN and MOLP implementations. On the other hand, from the point of view of recognition processing time, the SOM model is relatively slow. A general conclusion is that the basic MLP approach is not suitable for efficient speech recognition. The static MOLP with non-constructive learning gives considerably better performance than an equivalent MLP and has a very similar performance to that of the TDNN.

Among the issues for future research are the extension of the MOLP principles to the TDNN architecture and an evaluation of these neural techniques when the size of the speech database is increased and their integration in particular application environments.

References

[Beau93] Beaugé L., Durand S., Alexandre F. Plausible self-organizing maps for speech recognition. In: Artificial Neural Nets and Genetic Algorithms, Albrecht R.F., Reeves C.R., Steele N.C. (eds.), (Proceedings of the International Conference in Innsbruck, 1993), Springer, Wien, 1993, pp. 221-226.

[Ben89] Bengio Y., Cardin R., de Mori R., Merlo E. Programable execution of multilayered networks for automatic speech recognition. *Communications of the ACM*, 32, 1989, pp. 195-199.

[Bott90] Bottou L., Fogelman S.F., Blanchet P., Lienard, J.S. Speaker-independent isolated digit recognition: multilayer perceptrons vs. dynamic warping. *Neural Network*, 3, 1990, pp. 453-465.

[Elm90] Elman J.L. Finding structure in time. *Cognitive Science*, 14, 1990, pp. 179-211.

[Frei93] Freisleben B., Bohn C.A. Speaker-independent word recognition with backpropagation networks. In: Artificial Neural Nets and Genetic Algorithms, Albrecht R.F., Reeves C.R., Steele N.C. (eds.), (Proceedings of the International

Conference in Innsbruck, 1993), Springer, Wien, 1993, pp. 243-248.

[Fu94] Fu L.M. Neural networks in computer intelligence. McGraw-Hill, New York, 1994.

[Gar93] Santos-García G. The Hopfield and Hamming networks applied to the automatic speech recognition of the five Spanish vowels. In: Artificial Neural Nets and Genetic Algorithms, Albrecht R.F., Reeves C.R., Steele N.C. (eds.), (Proceedings of the International Conference in Innsbruck, 1993), Springer, Wien, 1993, pp. 235-242.

[Koh88] Kohonen T. The "neural" phonetic typewriter. *IEEE Computer Magazine*, March, 1988, pp. 11-22.

[Koh89] Kohonen T. Self–organization and associative memory. Series in Information Sciences, vol. 61, Springer, Berlin, 1989.

[Owe96] Owens F.J., Zheng G.H., Irvine D.A. A multi-output-layer perceptron. *Neural Computing & Applications*, 4, 1996, pp. 10-20.

[Tatt90] Tattersal G.D., Linford P.W., Linggard R. Neural arrays for speech recognition. In: Speech and Language Processing, Wheddon C., Linggard R. (eds.), Chapman and Hall, London, 1990, pp. 245-290.

[Tom89] Tom M.D., Tenorio M.F. A spatio-temporal pattern recognition approach to word recognition. In: Proceedings of IJCNN, Washington DC, vol. I, 1989, pp. 351-355.

[Vee95] Veelenturf, L.P.J. Analysis and applications of artificial neural networks. Prentice Hall, London, 1995.

[Wai89] Waibel A., Hanazawa T., Hinton G., Shikano K., Lang K. Phonem recognition using time-delay neural networks. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37, 1989, pp. 328-339.

[Zhe93] Zheng G.H., Owens F.J. A multi-layer neural network with a multi-output layer. In: Proc. Int. Conf. Neural Networks and Signal Processing, Guangzhou, China, 1993, pp. 46-50.

[Zhe96] Zheng G. H., Design and Evaluation of a Multi-Output Layer Perceptron, DPhil Thesis, University of Ulster, 1996.