# A HIGH-ORDER TEMPORAL NEURAL NETWORK FOR WORD RECOGNITION

Q.J. Zhang , Fang Wang and M.S. Nakhla

Dept. of Electronics
Carleton University
Ottawa, Canada, K1S 5B6

## ABSTRACT

An important yet challenging task for neural network based speech recognizers is the effective processing of temporal information in speech signals. A high-order fully recurrent neural network is developed to effectively handle the sequential nature of speech signals and to accommodate both temporal and spectral variations. The proposed neural network has 4 layers, namely, the input layer, self organizing map, fully recurrent hidden layer and output layer. The important characteristics of the hidden neurons and the output neurons are their high-order processing feature. A 2-stage unsupervised/supervised training method is developed. The solution from unsupervised training provides a good starting point for supervised training. The proposed neural network and the training method are applied to isolated word recognition using the TI20 data.

## 1. INTRODUCTION

Neural network methods have been very successful in many signal processing applications [1]. However the effectiveness of neural networks in speech recognition has been seriously affected by their inadequate capability to handle temporal variation and sequential information which are inherent in speech signals. This problem has received attention from many researchers and a number of techniques have been developed. The Time-Delay Neural Network (TDNN) [2], for example, is a very attractive architecture for learning the temporal relationships in acoustic events with reduced sensitivity to shifts in time. Recent advance in this approach is the Multi-State TDNN and the addition of a nonlinear time alignment procedure (DTW) to further improve the time-alignment capability of the model [3] [4] . Another popular approach is the hybrid system of neural networks and Hidden Markov Models (HMM), e.g., [5] [6]. In this case the neural network can be used, e.g., to approximate functions for computing acoustic parameters to be used as observations by HMM [6]. In these approaches, processing temporal information is partly or mainly done using mechanisms other than the neural network, i.e., using DTW in [3] and HMM in [5] [6]. A drastic and interesting solution is to specifically construct a neural network to virtually emulate the mechanism of hidden Markov Models and to implement Viterbi algorithms, e.g., [7] [8], fully taking advantage of the powers of HMM in speech recognition.

Techniques to improve the temporal processing capability of neural networks without explicit use of DTW or HMM concepts have also been studied. Among these approaches are, for example, the combined multilayer perceptron/self-organizing maps (SOM)[9] where two separate SOM are used to split the overall signal into first and second halves, the sequential competitive avalanche field where a constellation of activity develops over selected nodes by letting previously fired neurons decay slowly [10], and the use of recurrent neural networks which processes signals in both space and time domains [11] [12].

On the other hand, recent studies on high-order neural networks [13] [14] revealed its potential for processing temporal information in signals. The objective of this paper is to extend such concept to speech recognition and to allow a neural network to do both time alignment and sequential information processing in speech signals. The neural network proposed contains two fully recurrent layers of high-order neurons. The use of high-order neurons facilitates the modulation effects between different neuron responses in the network, and handles the temporal information more directly than standard first-order neurons. To speed up training, a two-stage unsupervised/supervised training method is developed. In the unsupervised training stage, assumptions about the network and the network parameters are employed. The solution from this stage is then used as a starting point for the next stage of

0-7803-2431-5/95 $4.00 © 1995 IEEE

training, i.e., supervised training. The technique is applied to word recognition using the TI 20 speech data.

## 2. ARCHITECTURE OF THE NEURAL NETWORK

The proposed neural network has 4 layers as shown in Fig. 1.

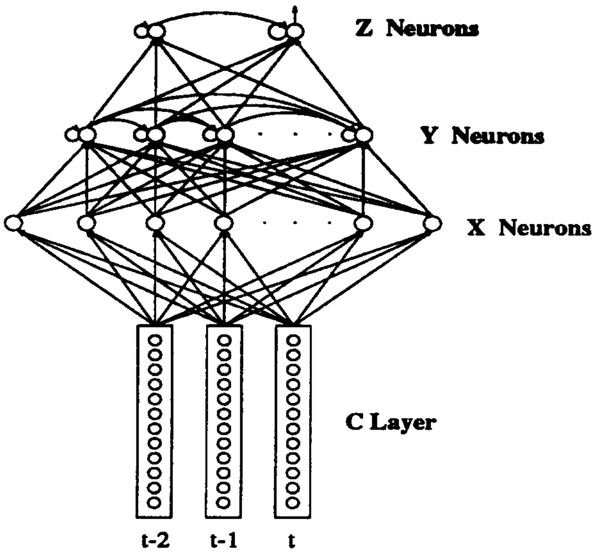

Figure 1: Architecture of the neural network

The input layer (*c* layer) has a buffer accepting 3 consecutive frames of feature parameters, e.g., cepstral parameters. The next layer (*x* layer) is a self organizing feature map with M neurons. The hidden layer (*y* layer) has N+1 fully connected recurrent neurons per word. The output layer (z layer) contains 2 recurrent neurons per word.

This neural network differs from typical recurrent networks in that neurons in the hidden and the output layers are high-order processing elements.

Let $x_i(t)$ represent the *ith* neuron in the feature map at time frame t, t=1, 2, ..., T, where T is the total number of frames of a speech utterance. $x_i(t) = 1$ or $0$. The output of neuron $k$ at the $y$ layer is

$$y_k(t) = f(\sigma_k(t)), k = 0, 1, 2, ..., N \tag{1}$$

$$\begin{aligned}\sigma_k(t) = {} & \lambda_k y_k(t-1) y_0(t-1) \\ & + \sum_{i=1}^{M} w_{ik} y_k(t-1) x_i(t) \\ & + \sum_{i=1}^{M} u_{ik} x_i(t) x_i(t-1) x_i(t-2) \\ & + \sum_{i=1}^{M} \sum_{j=1}^{N} v_{ijk} y_j(t-1) x_i(t) x_i(t-1) x_i(t-2)\end{aligned} \tag{2}$$

where $f$ is a sigmoid function, defined by

$$f(\sigma) = \frac{1}{1+e^{-\sigma}} \tag{3}$$

and $u_{ik}, v_{ijk}$ and $w_{ik}$ are weighting factors. The purpose for the $y$ layer is to encode the sequence of neuron firings in the feature map. The high-order term with coefficients $u_{ik}$ is only used for a special neuron $y_0(t)$ to detect the existence of 3 consecutive frames of signals causing the same $x$ neuron to fire. Using the high-order term with $v_{ijk}$, neuron $y_k(t)$ will fire only if 1): three consecutive frames of signals consistently encourage $y_k$ to fire, and 2): a sequentially correct $y_j$ is high in the previous time frame. If three consecutive frames of signals consistently encourage neurons other than $y_k$ to fire, the high order term with $\lambda_k$ will help to make $y_k$ decrease because of negative value of $\lambda_k$. Using the high order term with $w_{ik}$ time alignment can be achieved. If $y_k$ represents firing of $x_i$ and $x_i(t)$ is firing at time $t$, this term will help to keep $y_k(t)$ high under the condition that $y_k$ at time $t-1$ is high. The number of neurons in the hidden layer, i.e., N, is independent of the speech length. A conceptual interpretation, for N=24 as an example, could be that there are 4 speech units (e.g., phonemes) in a word model and in each unit there are 6 clusters of non-accidental variations in the cepstral space.

The output layer is described by

$$\begin{aligned}z_l(t) = {} & \mu_l z_l(t-1) + \sum_{k=1}^{N} \nu_{kl} y_k(t) \\ & + \sum_{k=1}^{N} \sum_{j=1}^{2} \xi_{klj} y_k(t) z_j(t-1), t \leq T, l = 1, 2\end{aligned} \tag{4}$$

where $\mu_l, \nu_{kl}$ and $\xi_{klj}$ are weighting factors. $z_1(t)$ does time integration to give the overall values in the $y$'s. $z_2(T)$ is the final criteria for recognition. The second-order term in (4) is used so that the word with the highest overall $y$ values are selected conditional upon the firing of some $y$ neurons which represent the speech units at the end of a word.

## 3. TWO-STAGE UNSUPERVISED/SUPERVISED TRAINING

To train the proposed neural net with a backpropagation based supervised training is extremely slow.

Table 1: PARAMETER CONSTRAINTS IN UNSUPERVISED TRAINING

| |
|---|
| for $k = 0, 1 \leq i \leq M, 1 \leq j \leq N$:<br>$u_{ik} = -1, v_{ijk} = w_{ik} = \lambda_k = 0$ |
| for $1 \leq k \leq N, 1 \leq i \leq M$ :<br>$u_{ik} = 0, \lambda_k = 1.$<br>if sequence from $y_j$ to $y_k$ is illegal,<br>then $v_{ijk} = 0$.<br>$y_k$ may represent the firing of<br>one or several $x_i$'s.<br>if $y_k$ not representing firing of $x_i$,<br>then $v_{ijk} = w_{ik} = 0$ .<br>if $y_k$ represents firing of $x_i$,<br>then $v_{ijk} = w_{ik} = 0$ . |
| for $1 \leq k \leq N, j = 1, 2$ :<br>$\mu_1 = 1, \mu_2 = \xi_{k1j} = \nu_{k2} = 0$ |

Table 2: $y$ HIDDEN LAYER SIZE FOR DIFFERENT WORDS

| word | N | word | N |
|---|---|---|---|
| one | 24 | go | 15 |
| two | 16 | no | 22 |
| three | 18 | enter | 20 |
| four | 19 | erase | 22 |
| five | 21 | rubout | 21 |
| six | 18 | repeat | 19 |
| seven | 21 | start | 16 |
| eight | 18 | stop | 18 |
| nine | 17 | help | 14 |
| zero | 19 | yes | 19 |

Here we exploit the neural net structure and its interpretation to create a more efficient 2-stage unsupervised/supervised training approach.

In **unsupervised training**, the sequence of neuron firings in the $x$ layer are extracted and a condensed version is encoded into $y$ layer. $u_{ik}, v_{ijk}$ and $w_{ik}$ are trained such that each $y$ neuron would represent a cluster of $x$ neurons and the sequence of the $x$ clusters that can fire consecutively is encoded into $v_{ijk}$. $\mu_l$ and $\nu_{kl}$ are predetermined structurally and $\xi_{klj}$ are obtained from statistics of the final $y$ neurons in speech utterances. Table 1 summarizes the choices of parameters in unsupervised training according to interpretations of the neural net for temporal processing. The solution from unsupervised training provides a good starting point for supervised training.

In **supervised training**, the objectives are 1): to emphasize the discrimination power of the neural network by simultaneously using both correct and wrong words in training, and 2): to free the neural network from some of the assumptions and choices imposed previously. Various neural net parameters are now allowed to change so the overall recognition accuracy is further improved.

## 4. EXAMPLES

We used the TI20 speech data which contains 10 digits and 10 command words such as Go, No, Start, Stop, etc. Signals were sampled at 12.5KHz with 12 bit resolution and pre-emphasized with a filter whose transform function is $1 - 0.95z^{-1}$. The waveform is then blocked into frames and multiplied by a Hamming window with length and shifts being 16ms and 10ms, respectively. From these smoothed speech samples 12 LPC-derived cepstral parameters per frame were computed. The cepstral parameters are the inputs of the neural network.

Different feature map sizes were experimented with. The size of the hidden y layer, i.e., N, varies according to words in the vocabulary. Table 2 lists such N values obtained from a feature map size of M=169. The speaker dependent recognition accuracy for the data was 95.5% for M=64, 99.2% for M=121 and 97.6% for M=169. The range of N typically are between 14 and 26 for various cases.

To show the processing of sequential information, consider example utterances of "one" and "no" which are similar if one of them is reversed in time [9]. Both signals were fed into the "one" network. The sequences of $y$ neurons encouraged to fire by the $x$ layer are ($y_4$ $y_7$ $y_9$ $y_{10}$ $y_{11}$ $y_{13}$ $y_{15}$ $y_{17}$) and ($y_4$ $y_{13}$ $y_{11}$ $y_{10}$ $y_8$ $y_7$ $y_6$ $y_3$ $y_{17}$) , respectively. The first sequence actually fired. However the second sequence of $y$ neurons were prevented from firing by the high-order modulation effect in (2) since sequentially the previous $y$ has not fired. So the final $z$ output is low, resulting in a rejection of the wrong word.

To show the ability of time alignment, we examine two example utterances of the word "repeat" with large temporal variations shown in Fig. 2. The sequences of neurons fired in the hidden layer of the "repeat" network are ($y_4$ $y_7$ $y_{10}$ $y_{11}$ $y_{13}$ $y_{15}$ $y_{16}$ $y_{17}$ $y_{20}$ $y_{21}$ $y_{22}$) and ($y_1$ $y_6$ $y_8$ $y_9$ $y_{11}$ $y_{13}$ $y_{15}$ $y_{16}$ $y_{19}$ $y_{20}$ $y_{21}$), which include cepstral variations between the utterances and the $y$ sequences are much less dependent on temporal distortions and speech lengths. All such utterances were correctly recognized.

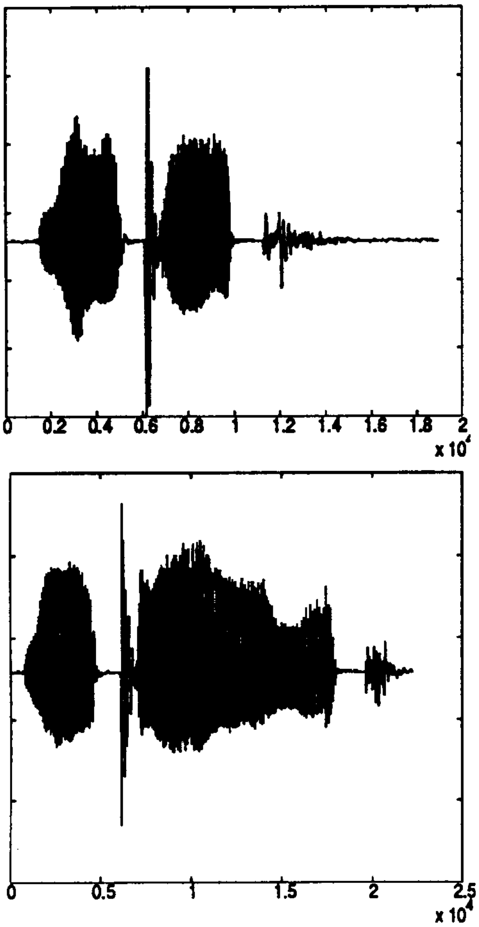

Figure 2: Two examples waveforms of "repeat" used to demonstrate time alignment

## 5. CONCLUSION

A high order fully recurrent neural network and a related training method have been proposed. High order neurons together with recurrency have been found to handle the sequential nature and temporal variations more directly and efficiently. A good recognition result of TI20 data has been achieved by using the proposed structure and training method.

## 6. REFERENCES

[1] S. Haykin, *Neural Networks, A Comprehensive Foundation*, New York, NY: IEEE Press, 1994.

[2] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K.J. Lang, "Phoneme recognition using time-delay neural networks", *IEEE Trans. Accust., Speech, Signal Processing*, vol. 37, pp. 328-339, 1989.

[3] H. Hild and A. Waibel, "Multi-speaker / speaker-independent architectures for the multi-state time delay neural network", *Proc. ICASSP*, vol. II, pp. 255-258, 1993.

[4] P. Haffner, M. Franzini and A. Waibel, "Integrating time alignment and neural networks for high performance continuous speech recognition", *Proc. ICASSP*, pp. 105-108, 1991.

[5] H.A. Bourlard and N. Morgan, *Connectionist Speech Recognition, A Hybrid Approach*, Boston, MA: Kluwer, 1994.

[6] Y. Bengio, R. De Mori, G. Flammia and R. Kompe, "Global optimization of a neural network-hidden Markov model hybrid", *IEEE Trans. Neural Networks*, vol. 3, pp. 252-259, 1992.

[7] R.P. Lippmann and B. Gold, "Neural-net classifiers useful for speech recognition", *Proc. ICNN*, vol. IV, pp. 417-425, 1987.

[8] L.T. Niles and H.F. Silverman, "Combining hidden Markov models and neural network classifiers", *Proc. ICASSP,(Albuquerque, NM)*, pp. 417-420, April, 1990.

[9] Z. Huang and A. Kuh, "A combined self-organizing feature map and multilayer perceptron for isolated word recognition", *IEEE Trans. Signal Processing*, vol. 40, pp. 2651-2657, 1992.

[10] J.A. Freeman and D.M. Skapura, *Neural Networks, Algorithms,Applications and Programmimg Techniques*, Reading, MA: Addison Wesley, 1992.

[11] G. Kuhn and R.L. Watrous, "Connected recognition with a recurrent network", *Speech Communication*, vol. 9, pp. 41-48, 1990.

[12] S.J. Lee, K.C. Kim, H. Yoon and J.W. Cho, "Application of fully recurrent neural networks for speech recognition" *Proc. ICASSP*, pp. 77-80, 1091.

[13] R.L. Watrous, "Speaker normalization and adaptation using second-order connectionist networks", *IEEE Trans. Neural Networks*, vol. 4, pp. 21-30, 1993.

[14] J. Heeb and L.A. Akers, "A temporal neural system", *Proc. Int. Symp. Circuits and Systems, (London, England)*, pp. 277-280, 1994.