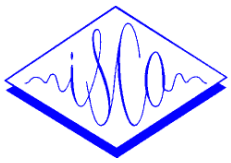

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

EXTENDED ELMAN'S RECURRENT NEURAL NETWORK FOR SYLLABLE RECOGNITION

Yong Duk Cho†, Ki Chul Kim, Hyun Soo Yoon, Seung Ryoul Maeng, and Jung Wan Cho

Deparment of Computer Science & Center for Artificial Intelligence Research
Koread Advanced Institute of Science and Technology
P.O.Box 150, Cheongryang, Seoul 130-650, Korea

Abstract

This paper describes an extended Elman's recurrent neural network adapted for speech recognition with input context buffers and analog target function. The input layer has context buffers to extract context sensitive features in the input. The analog target function in the output layer reflects the confidence level of the output for the current input in the context buffer. Speaker dependent recognition results for 10 syllables using cepstral coefficients show that the extended Elman's network is superior to the Elman's network as well as Multi-layer Perceptron. The recognition accuracy of the extended Elman's network is better than that of the cepstral distance measure and comparable to that of the weighted cepstral distance measure using dynamic time warping based template matching. Preliminary conclusion is that the input context buffers with time replicated scanning enhance the shift invariant capability of the recurrent neural network.

I. Introduction

Recently, various neural networks have been studied for temporal sequences such as speech signal. The successful neural networks for speech recognition should not only capture the temporally-distributed features, but also allow the temporal distortion that results in length variation. Conventional approaches for the sequential input use spatial metaphors for time dimension [1, 2, 3], that is, these approaches parallelize time axis by giving the network a spatial representation. However, this increases the network complexity. To recognize the speech signal, this type of neural networks should normalize the input length [2], or prepare a large number of input neurons by padding zeros for the unoccupied neurons with a buffer size of the maximum input length [3]. A more flexible model is preferred for speech recognition without preprocessing step and increase of the network complexity.

Recurrent neural networks can recognize time-varying sequences by using the recurrent connections that give the network memory [4, 5, 6]. Although the recurrent connections provide some capacity of the sequence recognition, it is too burdensome for them to memorize all the dynamicity in the speech signal with them only. In this paper, we tried to use both the recurrent connections and input buffer to enhance the capacity of the temporal memory for the variable length of speech sequences. We extended one of the recurrent neural networks that has fully recurrent connections among the hidden neurons, which is of the Elman's recurrent neural network. We extended the input layer of the Elman's network with the context buffers which are useful to extract context sensitive features in the input [7]. Generally used binary target function is replaced with an analog(ramp) target function to regulate the output level during the training phase. This reflects the confidence level of the output for the current input in the context buffer. We have evaluated the proposed network with 10 nasal consonant-vowel pair discrimination experiments in a speaker dependent mode. The improved results show the utility of the proposed extensions.

II. Extended Elman's Recurrent Neural Network

2.1 Network Architecture

Elman [4] proposed a neural network that has recurrent connections in hidden layer, and tried to catch the dynamicity in parsing a natural language. However, it seems hard to recognize spectral and temporal variances inherent in speech signal with the network, because it depends only on the recurrent connections for temporal integration. So, we extended the Elman's network. The input layer of the extended Elman's network consists of n ($n > 1$) context buffers instead of 1 in the Elman's to extract context sensitive features in the input pattern. Fig. 1 shows the topology of the extended Elman's recurrent neural network. The neurons between the adjacent layers are fully connected.

The input to neuron i at time τ, net_i^τ, is defined as a linear function of neurons j which are connected to neuron i as below:

† He has been with Samsung Electronics Corp. since Sep. 1990.

$$net_i^\tau = \begin{cases} \sum_j x_j^\tau w_{ij} + \sum_k x_k^{\tau-1} r_{ik}, & \text{if } i \in \textit{hidden neurons} \\ \sum_j w_{ij} x_j^\tau, & \text{if } i \in \textit{output neurons} \end{cases}$$

where w_{ij} is the feedforward connection strength from neuron j to i and r_{ik} is the recurrent connection strength from neuron k to i. The role of the recurrent connections at time τ is to integrate last sub-sequences which are already scanned from time 1 to τ-1. The activation value of neuron i at time τ, x_i^τ, is given below:

$$x_i^\tau = \frac{1}{1+e^{-net_i^\tau}}$$

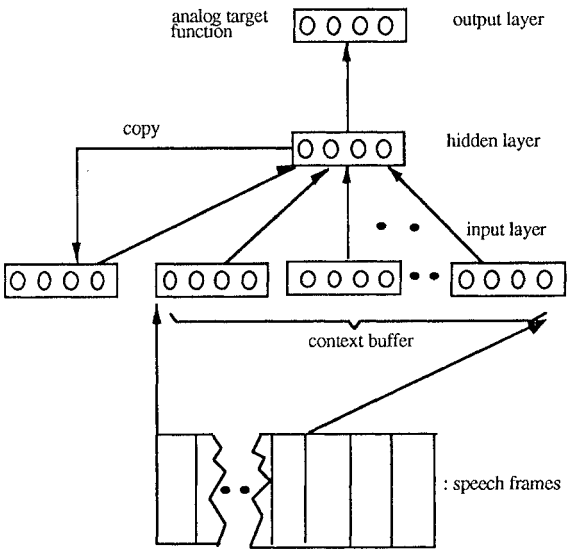

Figure 1. Extended Elman's recurrent neural network

2.2 Learning Formalism

We used the learning formalism for the recurrent neural network analyzed by Pineda [8]. The measure of the error in the network at time τ, E^τ, is defined as below:

$$E^\tau = \frac{1}{2}\sum_i (J_i^\tau)^2, \qquad \text{if } i \in \textit{output neurons}$$

where J_i^τ is given by

$$J_i^\tau = g_i^\tau - x_i^\tau$$

where g_i^τ is the desired output in output neuron i at time τ and x_i^τ is real output. An analog target function [6], g_i^τ, reflects the confidence level of the desired output value for the present input sub-sequence. As scanning a pattern, the confidence level of the output neuron being trained is increased from 0.5 to 1.0, otherwise decreased from 0.5 to 0.0 as below:

$$g_i^\tau = \begin{cases} 0.5(1.0+\frac{\tau}{T}), & \text{if } i = \textit{training output neuron} \\ 0.5(1.0-\frac{\tau}{T}), & \text{if } i \neq \textit{training output neuron} \end{cases}$$

where T is the constant for the slope of the target function. Back propagation learning procedure minimizes the error in the network with the gradient descent method in the weight space. Weight changes occur after every forward propagation. For the feed forward connections, the amount of weight changes from neuron j to i at time τ, Δw_{ij}^τ, is

$$\Delta w_{ij}^\tau = \eta y_i^\tau x_j^\tau$$

where η is a learning rate and y_i^τ is given as below:

$$y_i^\tau = \begin{cases} f'(net_i^\tau) J_i^\tau, & \text{if } i \in \textit{output neurons} \\ f'(net_i^\tau)(\sum_k w_{ki}^\tau + \sum_l r_{li}^\tau), & \text{if } i \in \textit{hidden neurons} \end{cases}$$

For the recurrent connections, the amount of weight changes from neuron j to i at time τ, Δr_{ij}^τ, is

$$\Delta r_i^\tau = \eta y_i^\tau x_i^{\tau-1}$$

To accelerate the learning speed, a momentum α is included in the weight changes for the feed forward and recurrent connection respectively as below:

$$\Delta w_{ij}^\tau = \eta y_i^\tau x_j^\tau + \alpha \Delta w_{ij}^{\tau-1}$$

$$\Delta r_{ij}^\tau = \eta y_i^\tau x_j^{\tau-1} + \alpha \Delta r_{ij}^{\tau-1}$$

2.3 Recognition Rule

Recognition phase employs a winner-take-all rule [9] which allows the network to keep the most highly activated neuron in the output layer. The activation values of the output neurons are accumulated with every forward propagation until the end of the input sequence. The decision rule for output neuron i is as follows.

$$o_i = \begin{cases} 1, & \text{if } \sum_\tau o_i^\tau = \max_j \sum_\tau o_j^\tau \\ 0, & \textit{otherwise} \end{cases}$$

where j is an output neuron.

III. Experiments and Discussion

3.1 Environments

The speech data consist of the nasal consonants /m/ and /n/ followed by /a, ∂, o, u, i/, which were uttered by four

male speakers three times. These were low-pass filtered up to 4.7 kHz with attenuation slope of 48 dB/octave, then digitized at 10 kHz with 12-bit quantization. Endpoints of speech were detected manually. To obtain the smoothed spectral feature vectors, the fourteenth order LPC analysis was done on a Hamming-windowed input samples after preemphasis. The window length is 20 ms long, and input samples of 10 ms duration were overlapped at a time. From these 14 LPC coefficients, cepstral coefficients were derived by recursive equations [10]. The length of frames for a syllable is variable because of the speaking rate difference between the speakers. The shortest syllable has 11 frames and the longest has 17. Thus the average frame length is 15.0 and standard deviation is 1.535.

In the network, the number of input neurons per context buffer is 14, which is the same as the dimension of LPC derived cepstral coefficients. We experimented the proposed network by varying the number of the context buffers in the input layer from 1 to 17. The number of output neurons is 10 which is equal to the number of syllables to be classified. The number of hidden neuron depends on the size of context buffers. The learning rate and momentum are 0.1 and 0.9, respectively. The initial strength of the weights is distributed uniformly and randomly between –0.5 and +0.5. The learning repeats until total-sum-of-square-error of the network reaches β * (17 – context_buffer(#) + 1) / 17 where β is a constant. In these experiments, β is fixed to 5.0. Learning phase used the syllables uttered twice, and test phase used the other utterance. Because of the small data set, circular evaluation was done yielding the performance of the network being measured from 12 repeated experiments. The simulation was done under MIPS-3260 computer.

3.2 The effect of the context buffer

The simulation was performed by varying the number of context buffers and hidden neurons. The number of hidden neurons is selected when the recognition rate is the best one in three times test as listed in Table 1. Fig. 2 shows the evaluation results. Remarkable one is that recognition performance shapes a graph with a unique maximal point. The network with 1 context buffer integrates temporally distributed information with only recurrent connections. The network with 17 context buffers does not use recurrent connections. In other networks with multiple context buffers, temporal integration is done by context buffers as well as recurrent connections. The performance increase according to the number of context buffers represents the cooperation of the context buffers and recurrent connections for temporal integration. However, excessive context buffers let the network be sensitive to the shifting of features in a pattern yielding to performance decrease. This is because the positions of the features for a pattern are fixed in the input layer. Although a windowed scanning with time replication can reduce the position dependency of the network, it depends on the size of the context buffers. A proper number of context buffers could enhance the shift invariance of the network to optimize the performance.

Table 1. Storage requirement according to the number of context buffers (The number of hidden neurons is selected when the recognition rate is the hightest.)

context buffers(#)	1	3	5	7	10	15	17
hidden neurons(#)	15	20	25	30	35	40	50
neurons(#)	39	72	105	138	185	206	298
connections(#)	585	1440	2625	4140	6475	10400	12400

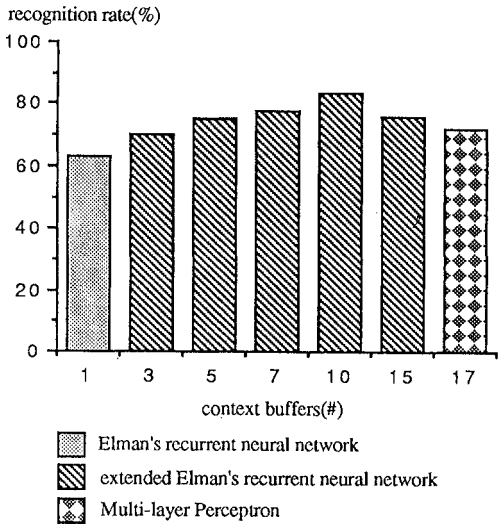

Figure 2. Performance evaluations according to the number of context buffers. If the number of context buffers is 1, extended Elman's recurrent neural network is equivalent to Elman's recurrent neural network, and if maximum(17), then Multi-layer Perceptron

3.3 Comparison with conventional template matching

We have compared the performance of the extended Elman's with the conventional template matching methods for cepstral coefficients, dynamic time warping (DTW) using the Euclidian cepstral distance measure and the index-weighted cepstral distance measure [11, 12]. The cepstral distance measure is defined as below:

$$dist = \sum_{i=1}^{14} (r_i - t_i)^2$$

where r_i is the i^{th} coefficient of the reference pattern r, and t_i is the i^{th} coefficient of test pattern t. With this measure, the recognition accuracy was 65.8 %, which is below than 83.3 % in the extended Elman's with 10 context buffers. The index-weighted cepstral distance measure [12] is known to show better performance than the cepstral distance measure, which is defined as bellow:

$$dist = \sum_{i=1}^{14} (i\,(r_i - t_i))^2$$

The recognition rate with the weighted cepstral distance measure was 82.5 %, which is comparable to that of the extended Elman's network.

3.4 Discussion

Conventionally, Multi-layer Perceptron(MLP) based neural networks have been widely used for pattern recognition, especially for speech recognition [2, 3]. However, the MLP is a static network, while the speech signal has dynamic characteristics. The extended Elman's network is a dynamic network, because it has feedback connections being natural for speech recognition. In the case of the maximal context buffers at the extended Elman's, the proposed network is equivalent to MLP [3] working with feedforward connections. The comparative recognition results for the extended Elman's, (pure) Elman's, MLP, cepstral distance measure, and weighted cepstral distance measure are shown in Table 2. The recognition rate of the extended Elman's is better than MLP as shown in Table 2. In addition, the extended Elman's requires less neurons and connections than MLP as shown in Table 1 (extended Elman's -> neurons : 185, connections : 6475 , context buffers : 10; MLP -> neurons : 289, connections : 12400, context buffer : 17). Both the recognition performance and space requirement of the extended Elman's are superior to those of MLP. The extended Elman's network is more natural for recognizing connected speech than MLP since it doesn't require length normalization procedure for the input. From Table 2, we can notice that the overall recognition rate is more or less low. This may be because of the lack of the training data set, we used only two repeated utterances for a syllable. It will be possible to improve the recognition performance by expanding the training set [13].

Table 2. Performance comparisions among extended Elman's recurrent neural network, Multi-layer Perceptron, Elman's recurrent neural network, cepstral distance measure and weighted cepstral distance measure

measure	recognition rate(%)
extended Elman's recurrent neural network	83.3
Multi-layer Perceptron	71.7
Elman's recurrent neural network	63.3
cepstral distance	67.8
weighted cepstral distance	82.5

IV. Conclusions

In this paper, we investigated an extended Elman's recurrent neural network for speech recognition, which has the fully recurrent connections between the hidden neurons. The input layer of the extended Elman's network adopts a context buffer which simultaneously scans speech frames more than two but not all, and the output layer uses an analog target function. The syllable recognition results show that the proposed neural network is superior both to the MLP and the Elman's recurrent neural network. Comparing the performance with the DTW based template matching, the recognition rate of the extended Elman's is superior to that of the cepstral distance measure and comparable to that of the weighted cepstral distance measure. We conclude that the recurrent connections and the context buffer could be used cooperatively to enhance the discrimination of the complex dynamic property in the speech signal. The extended Elman's recurrent neural network is a promising model for speech recognition, because it dose not need a complex segmentation procedure being extended for connected speech recognition with some decision mechanism when to fire the highly activated output neuron.

References

[1] T. J. Sejnowski and C.R. Rosenberg, "NETtalk: A Parallel Network that Learns to Read Aloud," *TR. JHU/EECS-86/01*, The Johns Hopkins University Electrical Engineering and Computer Science, 1986.

[2] D. J. Burr, "Experiments on Neural Net Recognition of Spoken and Written Text," *IEEE Trans. ASSP*, Vol. 36, No. 7, pp. 1162-1168, July 1988.

[3] R. K. Moore and S. M. Peeling, "Minimally Distinct Word-pair Discrimination using a Back-propagation Network," *Computer Speech and Language*, Vol. 3, No. 2, pp. 119-131, Apr. 1989.

[4] J. L. Elman, "Finding Structure in Time", *CRL-TR-8801*, University of Califonia, San Diege, 1988.

[5] M. I. Jordan, "Serial Order: A Parallel Distributed Processing Approach," *Institute for Cognitive Science Report 8604*, University of Califonia, San Diego, 1986.

[6] R. L. Watrous and L. Shastri, "Learning Phonetic Features using Connectionist Networks: An Experiment in Speech recognition," *TR. MS-CIS-86-78*, University of Pennsylvania, 1986.

[7] H. Boulard and C. J. Wellekens, "Speech Dynamics and Recurrent Neural Networks," *Proc. IEEE Int. Conf. ASSP*, pp. 33-36, Glasgow, 1989.

[8] P. J. Pineda, "Generalization of Back-propagation to Recurrent Neural Network," *Physical Review Letters*, Vol. 18, pp. 2229-2232, Nov. 1987.

[9] J. A. Feldmann and D. H. Ballard, "Connectionist Models and Their Properties," *Cognitive Science*, Vol. 6, pp. 205-254, 1982.

[10] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[11] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Trans. ASSP*, Vol. 26, No. 1, pp. 43-49, Feb. 1978.

[12] B. A. Hanson and H. Wakita, "Spectral Slope Based Distortion Measures for All-pole Models of Speech," *Proc. IEEE Int. Conf. ASSP*, Tokyo, pp. 757-760, 1986.

[13] T. Matsuoka, H. Hamada and R. Nakatsu, "Syllable Recognition using Integrated Neural Networks," *Int. Joint Conf. on Neural Networks*, Washington, pp. I. 251-258, 1989.