TDNN-Based Speech Recognition for Dental Unit Control, Part I: Speaker Persian words Recognition

Mohammad Bagher Menhaj, Majid Namnabat and Mohammad Ahadi
Electrical Eng. Department - Amirkabir University of Technology
No. 424 Hafez Ave. Tehran 15914 Iran
Phone: 009821-6466009, Fax: 009821-6406469
email: menhaj@cic.aku.ac.ir

ABSTRACT: Advanced dental units can lead to improved hygiene in dentistry clinics. For this reason, a new system for the control of such instruments is under development by the authors. In the first stage, several speech recognition techniques were considered and Time-Delay Neural Network (TDNN) approach was found superior regarding to both recognition speed and accuracy issues. A simulation of such a recognition system was carried out on speaker isolated Persian words recognition problem and having tested on 10 speakers, almost acceptable results were achieved. The implementation and further developments in the recognition algorithm will be reported in near future.

KEYWORDS: TDNN Nets, Speech Recognition

INTRODUCTION

To improve hygiene in dentistry clinics advanced dental units are needed to develop. To do so, a new system for controlling such instruments is under development and speaker isolated Persian words recognition problem was our main concern in the first stage of our project. Although several neural network structures have been used in speech recognition tasks since 1987 [Morgan, (1991)], their major problem is found to be their weakness in modeling the dynamics of speech. Time Delay-line Neural Networks has been introduced to overcome this shortfall by applying a few changes to the basic structure of MLPs and has been successfully used for such tasks [Waibe, (1988)],[Benani, (1991)],[Waibel, (1990)],[Waibel, (1989)]. In this paper, several speech recognition techniques were considered and Time-Delay Neural Network (TDNN) approach was found superior regarding to both issues of recognition speed and accuracy.

PROBLEM STATEMENT

As mentioned above the speech recognition problem, which we are dealing with, is recognition of speaker isolated Persian words listed below. Table 1 lists the 12 most usual words in relation to the functions of the dental unit were chosen to constitute the recognition task. The speech used during these experiments can be considered as almost clean speech (S/N ratios of around 25 dB). However, noisier speech should be considered for the second stage to cope with real conditions.

FEATURE EXTRACTION

The block diagram of the feature extraction unit utilized throughout this work is shown in Figure 1. The speech signal is input to a sound blaster card via a microphone and sampled and converted to digital after being low-pass filtered. Sampling rates of 11.025 KHz and 16.0 KHz and both 8 bit and 16 bit A/D conversions were used in different stages of this work. An endpoint detection algorithm, which utilizes energy and zero-crossings is then used to find the starting

Table 1. The 12 most usual words

1- Bala	2- Paein	3- Chap	4- Raast	5- Livan	6- Kerashoovar
7- Cheragh	8- Tanzim	9- Kaafi	10- Scop	11- Makandeh	12- Kevitron

point of the utterance[Rabiner, 1975]. The word plus its trailing silence, totaling a length of one second is then cut from the input speech.

A filter ($H(z) = 1 - 0.9z^{-1}$) with a pole at z=0.9 was applied in the first stage. Next, each utterance was blocked into frames of 30 msecs each, with an overlap of 15 msecs between adjacent frames. A total of 61 frames were used for each word in this case, resulting in a word of size 956 msecs with a total sample count of 10540 per word. However, for the next experiments, the frame size was changed to 20 msecs with a frame overlap of 10 milliseconds. The next block in Figure 1 performs the windowing. This consists of a Hamming window, applied to the frames of speech as follows.

$$\tilde{X}_m(k) = W(k).X_m(k)$$

where

$$W(k) = 0.54 - 0.46\cos(2pk/K).$$

An LPC analysis with an order of p=10 is carried out later. This consists of an autocorrelation analysis and the application of Durbin algorithm in order to find the LPC parameters. 10 Cepstral parameters are then calculated iteratively from the LPC parameters [Tohkura, (1987)]. A Juang lifter is then applied to the cepstral parameters and 10 delta-cepstral parameters are further aclculated and added to the results after filtering. The delta-energy parameter is then calculated using the energy of the "zero'th cepstral parameter"; this leads to an increase in the total number of parameters in the feature vector of each frame to 21. Finally, to avoid the neural network instability, the input vector should be normalized. To do so, two approaches might be taken. The first consists of dividing the feature vector elements to certain constant value, such as the number of samples per frame. The second is to use the log parameters.

THE TDNN MODEL

The TDNN is indeed a multilayer feedforward network, whose general structure is shown in Fig.2, in which the inputs of the each layer are buffered several time steps and then fully connected to the next layer. The TDNN used in this paper consists of an input layer, two hidden layers and one output layer. The input layer contains (21*61) nodes encoding the spectrum (i.e. each column represents features associated with a 30msec frame of the speech). The first hidden layer consists of 29 copies of 18 hidden units, the second hidden layer contains of 12 copies of 12 hidden neurons, and the output layer consists of 1 copy of 12 output units. The 12 outputs of the TDNN account for the recognition of 12 words. By referring to figure 2, one can show that the following equations describe the behavior of TDNN:

Forward path:

$$n_i^1(t+m\Delta) = \sum_{d=0}^{D_1=61} \sum_{j=1}^{N=21} W_{d,i,j}^1 P_{j,d+1}(t+m\Delta) + b_i^1$$

$$,1 \le i \le S_1 = 18, 0 \le m \le M - D_1$$

$$a_i^1(t+m\Delta) = F^1(n_i^1(t+m\Delta))$$

$$A^1 = \left\{\underline{a}^1(t), \underline{a}^1(t-\Delta); \cdots; \underline{a}^1(t-D_2\Delta)\right\},$$

$$n_i^2(t+m\Delta) = \sum_{d=0}^{D_2=29} \sum_{j=1}^{N2=18} W_{d,i,j}^2 A_{j,d+1}^1(t+m\Delta) + b_i^2$$

$$,1 \le i \le S_2 = 12, 0 \le m \le M_1 - D_2$$

$$a_i^2(t+m\Delta) = F^2(n_i^2(t+m\Delta))$$

$$A^2 = \left\{\underline{a}_1^2 \underline{a}_2^2; \cdots; \underline{a}_{12}^2\right\}$$

$$n_i^3 = \sum_{d=0}^{D_3=12} \sum_{j=1}^{N3=12} W_{d,i,j}^3 A_{j,d+1}^2 + b_i^3$$

$$,1 \le i \le S_3 = 12$$

$$a_i^3 = F^2(n_i^2)$$

Backward path & Correction Terms Computations:

$$\bar{d}^3 = -\dot{F}^3(\bar{n}_i^3) \cdot E$$

$$dW^3 = -a\bar{d}^3 (A^2)^T$$

$$\bar{d}_d^2 = -\dot{F}^2(\bar{n}_{d+1}^2)(w_d^3)^T \cdot \bar{d}^3$$

$$; 0 \le d \le D_3$$

$$dW_d^2 = -a\bar{d}_d^2 (A_{d+1}^1)^T$$

$$dW^2 = \left(\sum_{d=0}^{D_3} dW_d^2\right)/(D_3+1)$$

$$\bar{d}_{d_3,d_2}^1 = -\dot{F}^1(\bar{n}_{d_3+1,d_2+1}^1)(w_{d_3}^2)^T \cdot \bar{d}_{d_3}^2$$

$$; \quad 0 \le d_2 \le D_2 \quad ; \quad 0 \le d_3 \le D_3$$

$$dW_{d_3,d_2}^1 = -a\bar{d}_{d_3,d_2}^1 (P_{d_3+1,d_2+1}^1)^T$$

$$\mathrm{dW}^1 = \left(\sum_{\mathrm{d}_3=0}^{\mathrm{D}_3} \sum_{\mathrm{d}_2=0}^{\mathrm{D}_2} \mathrm{dW}_{\mathrm{d}_3,\mathrm{d}_2}^1\right) / ((\mathrm{D}_2+1) \times (\mathrm{D}_3+1))$$

SIMULATION RESULTS

The simulation of the neural network was performed within 2 distinct experiments, i.e. speaker dependent (SD) and speaker independent (SI). For the SD experiment, the training set consisted of 50 repetitions of each word uttered by a single speaker, while only 10 repetitions of the same speaker were used for the purpose of testing. In SI case, the training set consisted of 60 and the test set consisted of 10 speakers. Half of both sets, in this case, were consisted of female speakers. Each speaker in this case uttered each word 3 times. In both cases, an offline test using Matlab and an online test using C-language were performed. The best recognition rate obtained for SD case was 87% while for the SI case, a score of 85% was obtained, which are still unacceptable for a real-world application.

CONCLUSION AND FUTURE WORK

In this paper, a time-delay neural network-based system for recognition of a number of key words for control of a dental unit was introduced. Several issues are believed to have considerable impact on the results obtained. Large range of length of the uttered words (300msec.s to 900 msec.s) and regional accents of some of the speakers cooperated in this project are among these issues. In order to achieve the specified goals in the implementation of this algorithm in a modern dental unit, the following research options are under investigations by the authors:

1. Application of an appropriate time normalisation technique to the uttered words (e.g. DTW or a similar technique).
2. Application of hybrid neural structures such as Kohonen/TDNN hybrid.
3. Optimizing network parameters.
4. Application of a speaker adaptation algorithm to improve the recognition rate for the specific user.

REFERENCES

Benani, Y. et al., 1991, "Validation of neural net. architectures on speech recognition tasks", ICASSP.

Morgan, D.P. and C.L. Schofield, 1991,"Neural Networks and Speech Processing", Kluwer Academic Publishers.

Rabiner, L.R., M.R. Sambur, 1875, "An algorithm for determining the endpoints of isolated utterances", Bell System Tech. J., vol.54, No.2.

Tohkura, Y., 1987, ”A weighted cepstral distance measure for speech recognition”, IEEE Trans. ASSP, vol.35, No. 10.

Waibel, A. et al., 1988, "Phoneme Recognition: Neural networks vs. hidden Markov models", in proc. ICASSP, New York, pp107-110.

Waibel, A. et al., 1990, "A time delay neural network architecture for isolated word recognition", Neural Networks, vol.3.

Waibel, A. et al., 1989, "Phoneme recognition using time delay neural networks", IEEE Trans. ASSP, vol.37, No.3.

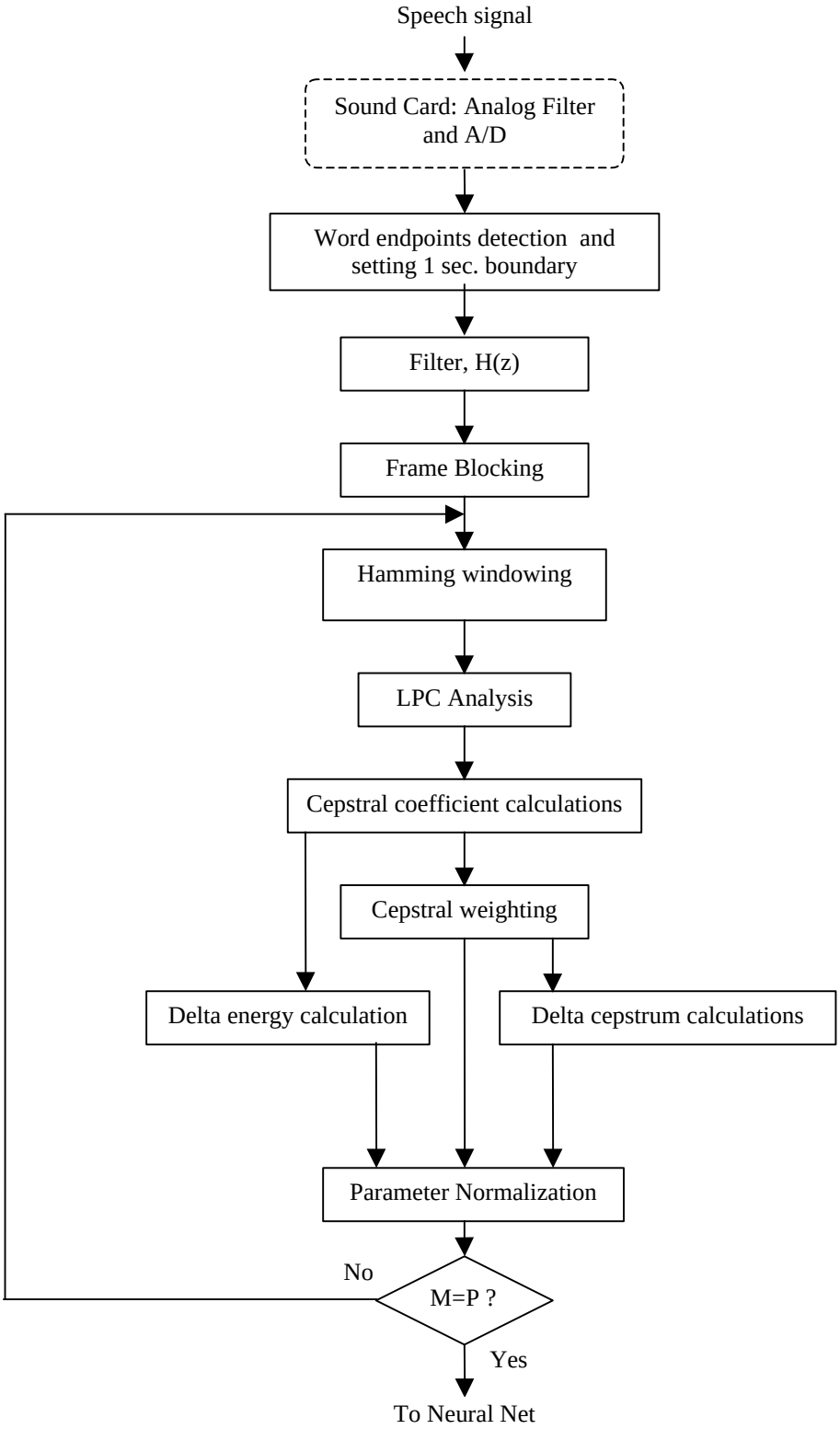
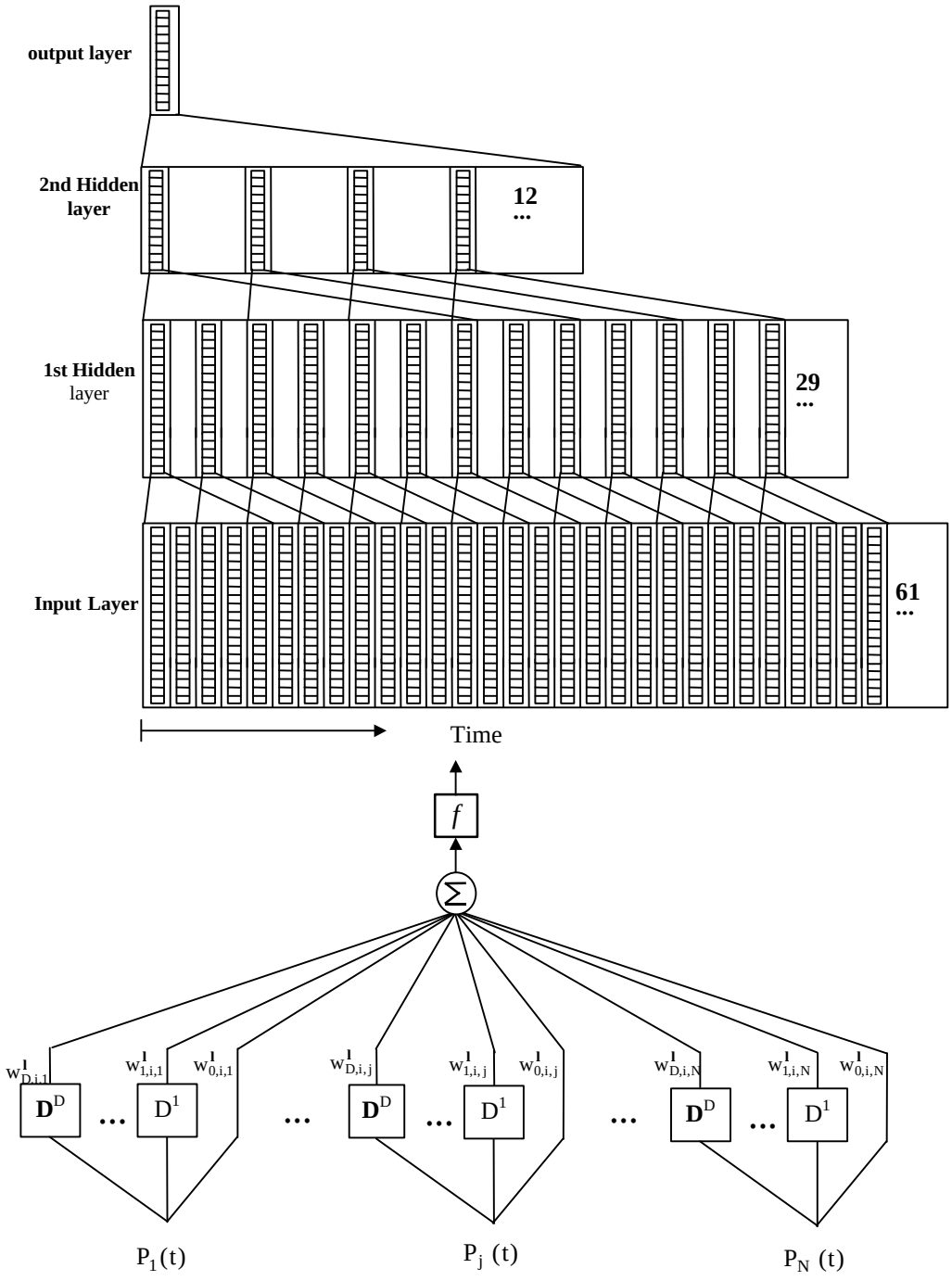


Figure 1. The feature extraction block diagram



(a)- The Structure Of each neuron in TDNNN

Fig. 2. The Designed TDNN