SPEECH ENHANCEMENT USING TIME DELAY NEURAL NETWORKS

M. I. Dawson and S. Sridharan

Signal Processing Research Centre
Queensland University of Technology

ABSTRACT - The use of Time Delay Neural Networks (TDNNs) for removing additive noise from continuous speech is described. Mel scaled frequency coefficients are used to parametise the noisy speech which is input to the network. The network is trained to extract the speech signal from the noise using the gradient back propagation algorithm. Preliminary results are presented.

INTRODUCTION

Neural networks have been used successfully in the past for noise reduction in speech signals. Of note are the papers of Tamura, (Tamura, 1988) (Tamura, 1990) and Sorenson (Sorenson, 1991) where mutli-layer neural nets were used with varying forms of preprocessing of the noisy speech signal. The encouraging results of Waibel's study into using TDNNs for phoneme recognition coupled with the strong spatio-temporal mapping features of this type of network lead to the choice of TDNNs as a good candidate for removing noise from speech.

This paper presents the preliminary results of an investigation into the use of a TDNN for removing noise from speech. The paper is divided into three sections. The first describes the TDNN and its network equations. The second section outlines the testing procedure and presents the results. The final section presents a summary and discusses directions for further research.

THE TDNN ARCHITECTURE

Waibel and co-researchers first introduced the Time Delay Neural Network in their 1989 paper (Waibel 1989b). Their TDNN was described as a translation invariant back-propagation network that performed better than sophisticated continuous acoustic parameter Hidden Markov Models on noisy speech recognition tasks. The temporal structure of the TDNN was found to well model the time invariant characteristics of speech. A succession of papers then followed, (Waibel, 1989a) (Lang, 1990) (Hampshire, 1990), highlighting various useful properties of the TDNN in which they obtained very good speech recognition results.

There is a significant difference in the way that Waibel used the TDNN for speech recognition tasks and the way that we have implemented it for additive noise removal from speech. For speech recognition the TDNN is set up as a classifier. That is, the output of the network is in the form of a classification of phoneme type when excited by input speech. In our network the output is in the same form as the input speech parameters, and as such, it is operating as a form of adaptive filter during training.

A spatio-temporal diagram of a small TDNN is shown in Figure 1. There are five input speech frames consisting of 16 Mel scaled frequency coefficients with a hidden layer of three frames of 16 elements and an output layer of one frame of 16 elements. Each frame in the hidden layer is fully connected to three consecutive frames in the input layer with a time shift of one frame in the input layer for each frame in the hidden layer. The output layer is fully connected to all three hidden layer frames. Thus the first digits in the term "5d2-3d2-1" refer to the number of frames per layer. The "d2" terms refer to the number of delays per layer (i.e. the current frame plus two delayed frames).

One problem in working with TDNNs is the visualisation of the active network parameters during training. Being able to observe the connection strengths between each processing element in the network is crucial in determining the learning abilities of the network. With 3072 weights in this relatively small network we are presented with the rather difficult problem of displaying these weight values. The

TDNN in Figure 1 displays the element output values as a square of varying colour and size. Element outputs range from -1.0 to +1.0. Positive values are displayed in black while negative values are in grey. This display does not give an indication of the weight values, however, we are experimenting with a feature display that plots the weight values in much the same way as a spectrogram. The important consideration here is to highlight the information content of the weight value magnitudes as well as the temporal structure as it pertains to the input speech data.

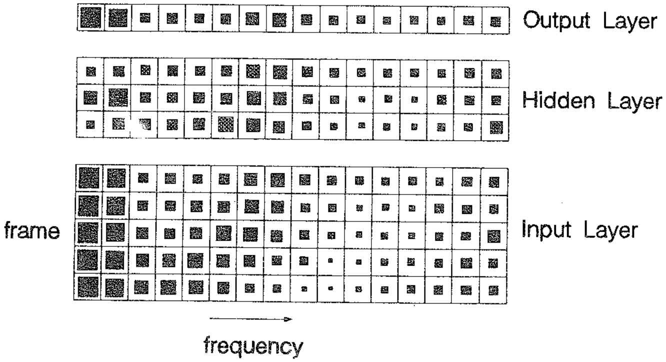

Figure 1. Spectro-Temporal Diagram of a 5d2-3d2-1 TDNN.

The TDNN architecture is trained using the gradient back propagation method which may be summarised as follows. The weights are first initialised to small random values. Experimentation with the efficacy of the network during training has yielded values between +/- 0.0625. The input pattern $\xi^{\mu}_{f_i k}$ is then applied to the input layer of the network so that

$$V^{0}_{f_i i} = \xi^{\mu}_{f_i i}, \qquad \text{for all } i.$$

Where f_i refers to the frame number and i refers to the processing element within that frame. There are P input patterns of class μ. The net values are propagated forward according to the formula

$$V^{m}_{f_i i} = g(h^{m}_{f_i}) = g\left(\sum_{f_i, i} \sum_{f_j, j} w^{m}_{f_i i f_j j} V^{m-1}_{f_j j} \right), \quad \textit{for all } f_i, i, f_j, j, m.$$

Where the function *g()* is the sigmoidal function.The *i*'s and the *j*'s refer to successive layers of the network. The deltas for the output layer are then calculated using

$$\delta^{M}_{f_i i} = g'\left(h^{M}_{f_i i}\right)\left(\zeta^{\mu}_{f_i i} - V^{M}_{f_i i}\right)$$

That is, each delta value for the output layer (layer *M*) is calculated by comparing the actual and desired network outputs and multiplying the difference by the differential of the processing element output. The errors are then propagated backwards through the network using

$$\delta_{f_ji}^{m-1} = g'\left(h_{f_ji}^{m-1}\right)\sum_{f_jj} w_{f_jjf_ji}^{m} \delta_{f_jj}^{m}, \quad for\ m = M, M-1, \ldots, 2$$

The weights are updated according to the rule

$$\Delta w_{f_jif_jj}^{m} = \eta \delta_{f_ji}^{m}(t) V_{f_jj}^{m-1} + \alpha \delta_{f_ji}^{m}(t-1)$$

where neta is the learning rate and alpha is the momentum term. In this equation, the index t is used loosely to refer to successive training epochs. It is important to note that this training procedure is applied to the network separately for each time shift. An entire copy of the network is stored for each time shift, and the resultant weight values are taken as the average of all time shifted images.

RESULTS

The TDNN was used to filter additive white noise from the five passages of speech shown below:

"Hayed, Heed, Hod, Hoed"
"We were away a year ago"
"I know when my lawyer is due"
"Every salt breeze comes from the sea"
"I was stunned by the beauty of the view"

The input speech was sampled at a frequency of 12 kHz with 16 bit resolution. Anti-alias filtering was performed inherently by the over-sampling analog-to-digital converters on board an Ariel DSP-32C installed in a 486 PC host. The data were then organised into frames of 256 samples with a 50% overlap and each multiplied by a Hanning window function. Pseudo-random white noise was added to the speech at varying signal to noise ratios and then 16 Mel scaled frequency coefficients were extracted from each frame of the data. The noisy speech coefficients were then used as the input to the TDNN with the clean speech coefficients used as the desired training set.

The five sentences were sampled and concatenated to form a training database of 1400 frames of coefficients. Two different modes of training were employed. The first was to present the entire database to the network at the one time. Results for signal to noise ratios of 10dB, 0dB, and -3dB are shown below in Figure 2. The second was to present a small portion of the database to the network and train it until convergence was reached, as done in (Waibel, 1989b). The training set was then doubled and training commenced again until convergence was reached. This pattern was continued until the whole database had been presented to the network. The results for a signal to noise ratio of 10dB are shown below in Figure 3. As can be seen, the incremental training method allows a much smaller mean squared error between the actual and desired outputs as well as faster training than the full training method.

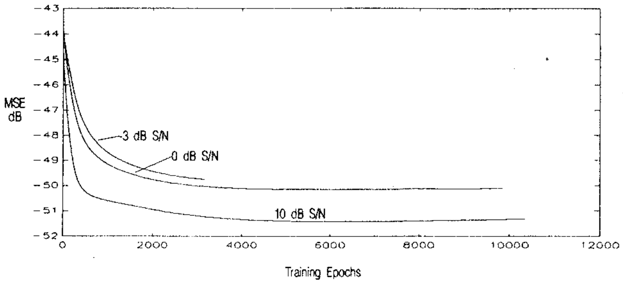

Figure 2. TDNN training using full data set - S/N ratios of 10, 0, and -3 dB.

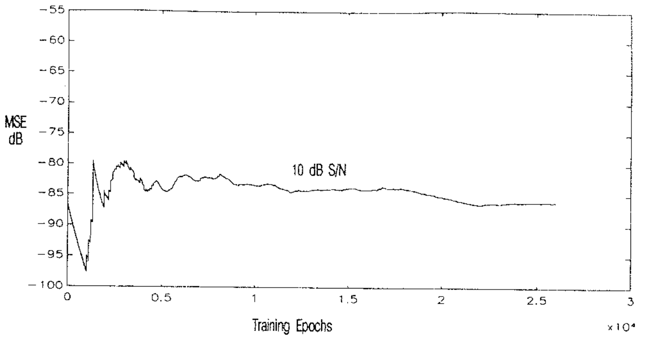

Figure 3. TDNN training using incremental data set - S/N ratio of 10 dB.

It was found that the sequential training method allowed the network to learn localised features of the database with greater accuracy. This method allows the network to incrementally improve its generalisation of the necessary features required to extract the speech from the noise. Simulations using the entire database as the training set from the start showed that the network learned very slowly and achieved nowhere near the levels of error minimisation than that obtained by incremental training.

CONCLUSIONS

A TDNN based noise removal architecture has been described. Preliminary results of the network's ability to filter white noise from speech are promising. Future investigations will address the ability of the network to generalise for speech not included in the training database. Further results will be presented at the conference.

REFERENCES

Hampshire, J.B. Waibel, A.H. (1990) *A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks,* IEEE Transactions on Neural Networks, Vol. 1. No. 2, June 1990, pp. 216-228.

Lang, K.J. Waibel, A.H. & Hinton, G.E. (1990) *A Time-Delay Neural Network Architecture for Isolated Word Recognition,* Neural Networks, Vol. 3, pp. 23-43.

Sorenson, H.B.D (1991) *A Cepstral Noise Reduction Multi-Layer Neural Network,* International Conference on Acoustics, Speech, and Signal Processing, ICASSP91, pp. 933-936.

Tamura, S. & Waibel, A.H. (1988) *Noise Reduction Using Connectionist Models,* International Conference on Acoustics, Speech, and Signal Processing, ICASSP88, pp. 553-556.

Tamura, S. & Nakamura, M. (1990) *Improvements to The Noise Reduction Neural Network,* International Conference on Acoustics, Speech, and Signal Processing, ICASSP90, pp. 825-828.

Waibel, A.H. Sawai, H. & Shikano, K. (1989a) *Modularity and Scaling in Large Phonemic Neural Networks,* IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37. No. 12, December 1989, pp. 1888-1898.

Waibel, A.H. Hanazawa, T. Hinton, G.E. Shikano, K. & Lang, K.J. (1989b) *Phoneme Recognition Using Time-Delay Neural Networks,* IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37. No. 3, March 1989, pp. 1888-1898.

AN ARTIFICIAL NEURAL NETWORK FOR AUTOMATIC SPOKEN LANGUAGE RECOGNITION

A.R. Dowling, R.W. King and J.P. Vonwiller

Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney

ABSTRACT - this paper describes an initial study into the use of an artificial neural network for discriminating between spoken Australian English, Mandarin Chinese and Lebanese Arabic. Seventeen features derived from the speech waveform and its spectrum were applied to the input of the network. The choice of these features was made from considerations of the known phonological and morphological differences between the languages. Language classification was better than 94% correct using 30-second samples of speech. The sensitivity of classification to each of the features, and to shorter speech samples are examined.

INTRODUCTION

Automatic recognition of the language being spoken by an individual may be of considerable practical value in future speech response and recognition systems, as well as having intrinsic scientific interest. The problem has been investigated with some success, using Hidden Markov models coupled with pitch contour analysis, to discriminate between English, Spanish, Hindi and Mandarin Chinese (Savic et al. 1991). In this paper we describe an initial study into using an artificial neural network (ANN) to discriminate between spoken Australian accented English, Mandarin Chinese and Lebanese Arabic. These three languages and dialects were chosen as major representatives of three language families, and are also well represented in the University community, so simplifying speech data collection.

The approach taken attempts to exploit the known major phonological and morphological differences between the languages. In essence, we expect that specific features in Arabic (such as the large proportion of pharyngeal and laryngeal sounds), and Mandarin (as a tonal language, with a highly constrained short syllable structure) would positively identify them. English would be identified by an absence of such specific features. For an automatic spoken language recognizer these discriminating features have, naturally, to be realised from the acoustic speech signal.

A practical recognition system should be designed to operate on an appropriately small feature set and on a minimum duration speech sample. The paper describes the choice of seventeen potentially discriminating features from the speech signal. These features are derived from 30-second samples of speech, and are used to train the multilayer perceptron (MLP) artificial neural network. The paper describes an examination of the sensitivity of language recognition to the members of the feature set, and the reduction in accuracy of recognition as the test sample duration of is reduced to 5 seconds.

DISCRIMINABLE FEATURES OF THE CHOSEN LANGUAGES

The three languages chosen for this study are from three language families; by definition this should make them more discriminable than languages from one family. The chosen languages do indeed exhibit morphological and phonological differences, as discussed below.

Lebanese Arabic

Arabic, the most widely spoken Semitic language from the Afro-Asian family, has a classical written form as well as many colloquial dialects. Our study was confined to speakers of the Lebanese dialect, reading from standard text.

The major features of interest here us arise from Arabic's simple vowel system, the large number of fricative consonants, and the pharyngeal and laryngeal consonants. The latter make the sound of Arabic more gutteral than other languages.