NONLINEAR ECHO CANCELLATION USING A PARTIAL ADAPTIVE TIME DELAY NEURAL NETWORK

A. N. Birkett, R. A. Goubran
Department of Systems and Computer Engineering
Carleton University, 1125 Colonel By Drive
Ottawa, Canada, K1S 5B6
Tel: (613) 788-2600 ext. 5740, Fax: (613) 788-5727
e-mail: birkett@sce.carleton.ca

Abstract: System identification of a nonlinear loudspeaker/microphone acoustic system is necessary to achieve high acoustic echo cancellation in the handsfree telephony environments where the loudspeaker often operates at high volumes. In this paper, a partial adaptive process consisting of a small order tapped delay line neural network (TDNN) followed by a delayed Normalized Least Mean Squares (NLMS) adaptive filter is used to model a loudspeaker/microphone acoustic system. The TDNN models the first part of the acoustic impulse response (AIR) where most of the energy is contained and the delayed NLMS filter models the remaining echo. Experimental measurements confirm that a short length TDNN is capable of improved identification in an undermodelled system and that by extending this to the partial adaptive TDNN structure, the ERLE performance improves by 5.5 dB at high loudspeaker volumes when compared to a NLMS structure.

INTRODUCTION

In this paper, a partial adaptive process consisting of a tapped delay line feedforward neural network (TDNN) and normalized least mean squares (NLMS) structure are employed in an attempt to model loudspeaker nonlinearities at high volumes. The specific application here is improved steady state performance for acoustic echo cancellers in the handsfree environment using conference type speakerphones. Most of these consumer products employ inexpensive audio components which are susceptible to nonlinear distortion at low frequencies and high volumes. In this paper, the identification of the nonlinear loudspeaker/microphone system is considered. In a real environment however, the AEC structure must be capable of identifying and tracking not only the reflected signals from the room, i.e. its acoustic impulse response (AIR), but also of modelling the plastic enclosure vibrations and nonlinear loudspeaker response, as shown in Figure 1.

.

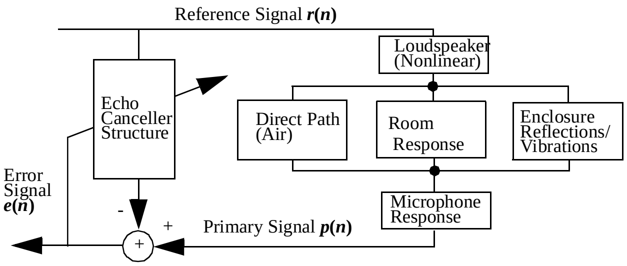

FIGURE 1. Acoustic Echo Canceller Structure. The AEC must identify not only the AIR but nonlinear and vibration effects as well.

Conventional AECs utilize a linear adaptive transversal filter to model the AIR and cancel the echo signal. At low volumes where nonlinearities are absent, this is a classical system identification problem whereby the adaptive filter adjusts its coefficients via the NLMS algorithm [4] to model the echo path, *H(z)* between the loudspeaker and the microphone so the system output, *e(n)* is minimized. The NLMS algorithm is the baseline by which performance of alternative models is measured but it is incapable of reducing nonlinear distortion. A measure of the AEC performance is the Echo Return Loss Enhancement (ERLE) which is defined as [5];

$$ERLE(dB) = \lim_{N \to \infty}\left[10\log\frac{E[p^2(n)]}{E[e^2(n)]}\right] \cong 10\log\left[\frac{\sigma^2{}_p}{\sigma^2{}_e}\right] \qquad (1)$$

where $\sigma^2{}_p$ and $\sigma^2{}_e$ refer to the variances of the primary and error signals respectively and E is the statistical expectation operator.

Limitations of AEC Performance

1) *TIP/TP Ratio*: One of the limitations of AECs is undermodelling of the AIR. As shown in [5] the achievable ERLE is determined in part by the degree of undermodelling of the unknown system. The results show that the achievable ERLE is determined by the Total Impulse Power to Tail Power (TIP/TP) ratio, defined as;

$$\frac{TIP}{TP} = 10\log\left[\frac{\sum_{i=0}^{M-1} h^2(i)}{\sum_{i=N}^{M-1} h^2(i)}\right] \qquad (2)$$

where h is an impulse of length M and N is the discrete point at which the “tail” is considered to start. The TIP/TP ratio is invaluable for determining the optimum number of AEC filter taps to use given a certain loudspeaker, microphone and enclosure. A TDNN can be used in place of a NLMS filter to identify a loudspeaker [7] and to see the effect on the TIP/TP ratio when operating at high loudspeaker volumes. The experimental results shown in Figure 6 indicate that the TDNN is capable of achieving a higher ERLE than the NLMS when in an undermodelled state, i.e. when the number of delays in the delay line is less than the AIR. This improvement in performance can be used in the partial adaptive structure (described below) to obtain improved performance at high volumes.

2) *Nonlinear Distortion:* The nonlinear parameters of a loudspeaker may be described by the force deflection characteristics of the loudspeaker cone suspension system and nonlinear flux density as described in references [1][2]and [3]. Suspension system nonlinearity manifests itself as soft clipping at the loudspeaker output and results in odd-order harmonics under large signal conditions. The nonlinear distortion consists mainly of cubic terms and can easily be 5 to 10 percent of the total output, especially when dealing with small loudspeakers that have low power ratings. Simulations and experimental results indicate that neural network models can identify this nonlinear distortion more effectively than linear adaptive structures.

3) *Enclosure Vibration*: Vibration is a serious problem that occurs under the same conditions as nonlinear distortion, namely at low frequencies and high volumes. It is important that this be addressed in a practice but is considered beyond the scope of this paper and will not be discussed further.

Figure 2 shows the general ERLE performance in a typical echo environment [5]. In a simulated experiment, the ERLE will follow the TIP/TP ratio very closely, however, in actual measurements, limitations such as room noise, vibration and loudspeaker nonlinearities will limit the achievable ERLE as indicated.

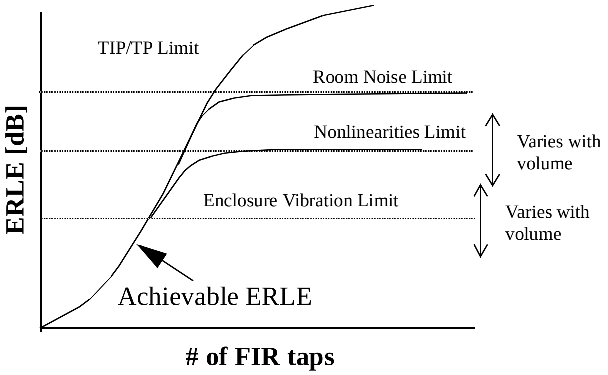

FIGURE 2. Achievable ERLE as a function of Physical Limitations. In the absenc of vibration, nonlinear distortion and room noise, the achievable ERLE is determined by the TIP/TP ratio.

NEURAL NETWORK MODELS

Tapped Delay Line Neural Network

A tapped delay line neural network previously presented in [6] is shown below in Figure 3. It can be used to perform a nonlinear system identification of the loudspeaker/microphone system. It consists of a tapped delay line input layer, two hidden layers which have a piecewise linear/sigmoidal activation function and a linear output node. The piecewise linear/sigmoidal activation function is linear below +/- 0.2 and then follows a squashed hyperbolic tangent sigmoid beyond this point such that the output is squashed between +/- 1.0. As shown in [6], this improves system identification at low volumes where the loudspeaker is essentially linear but also allows for the soft clipping effect observed at higher loudspeaker volumes. It should be noted that for activation levels less than the +/- 0.2 linear region, the gradient calculations are trivial and the filter complexity approaches that of the NLMS filter.

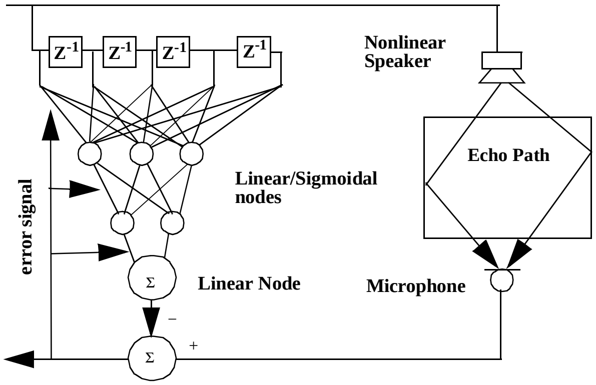

FIGURE 3. Tapped Delay Line Neural Network Adaptive Echo Canceller Structure (TDNN).

Partial Adaptive Model

The partial adaptive process utilizing a neural network preprocessor is shown in Figure 4. It consists of a low order TDNN to model the large part of the AIR and a NLMS filter to model the tail of the echo. A fixed delay line equivalent to the delay line length of the TDNN is inserted before the NLMS filter. A similar algorithm incorporating a multi-microphone linear NLMS structure is presented in [8]. In this paper however, the structure has been modified to incorporate a neural network as a nonlinear preprocessor.

Referring to Figure 4, the output $e_2(n)$ is given as;

$$e_2(n) = e_1(n) - y_2(n) = p(n) - y_1(n) - y_2(n) \tag{3}$$

where $p(n)$ is the microphone (primary) signal, $y_1(n)$ is the output of the TDNN and $y_2(n)$ is the output of the delayed NLMS filter. Expanding, we obtain;

$$e_2(n) = p(n) - y_1(n) - \sum_{i = N_1 + 1}^{N_2} w(n)x(n-i) \tag{4}$$

where $w(n)$ are the NLMS tap weights and $x(n)$ is the information vector, N_1 is the delay length of the TDNN section and N_2 is the total impulse length.

In the proposed structures there is no feedback hence the backpropagation algorithm [9] is employed to train the networks. A normalized step size [4] is employed during the training and tracking phase for both the NLMS and neural network sections. The stepsizes μ_{NLMS} and μ_{TDNN} are individually calculated and updated after each new sample is shifted into the tapped delay line.

The TDNN consists of 150 taps in the delay line, and 2 and 3 nodes respectively in the 1st and 2nd hidden layers. The NLMS section has 450 taps such that the total impulse response is 600 taps.

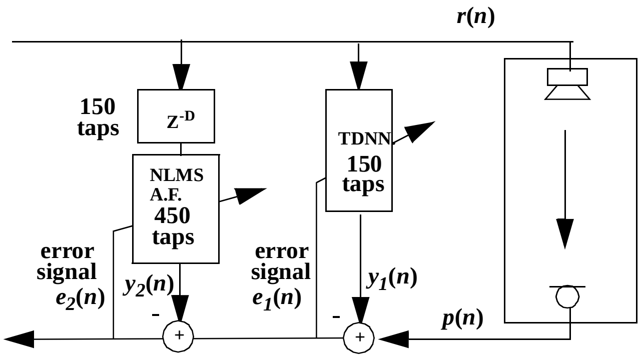

FIGURE 4. Partial adaptive structure utilizing a TDNN to cancel the first part of the AIR and a NLMS to cancel the tail portion. Signal $e_2(n)$ is the residual signal left after the echo has been cancelled.

COMPUTER SIMULATIONS

Simulations were performed using a computer generated white noise source as the reference signal, which was then filtered and convolved with an artificial room impulse function. The reference and primary files were then applied to the corresponding algorithms. For each run, the reference signal is distorted by adding both quadratic and cubic distortion according to the following equation;

$$y = \frac{ax + bx^2 + cx^3}{|a| + |b| + |c|} \tag{5}$$

where *a*, *b*, and *c* refer to the amplitude of the linear, quadratic and cubic factors, *x* is the input signal and *y* is the output signal level. The coefficients *b* and *c* were increased such that the distortion level increases relative to the primary signal level. The signal to distortion ratio is calculated by dividing the variance of the undistorted signal portion by the variance of the distorted signal portion. For each run, the algorithm was allowed to converge for 80000 samples and then a mean converged ERLE was obtained. The results shown below in Figure 5 indicate that the partial adaptive network outperforms the NLMS in high distortion environments, i.e. at low Primary/Distortion ratios.

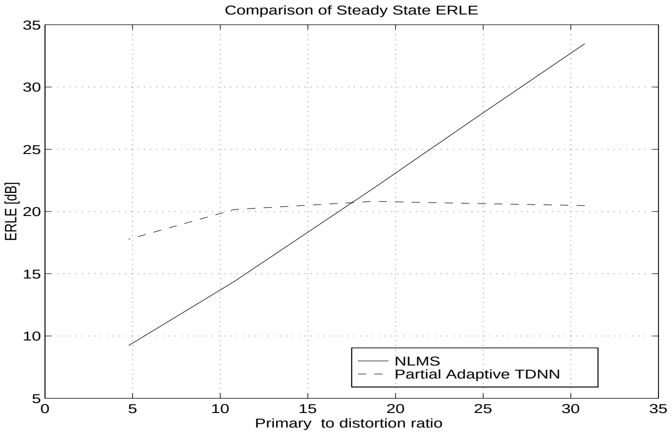

FIGURE 5. Simulation results show that the partial adaptive TDNN outperforms the NLMS in high distortion environments.

EXPERIMENTAL RESULTS

Experimental Setup

In order to remove the effects of vibration and room noise, the loudspeaker and microphone from a commercially available speakerphone were removed and placed in a standard baffle inside an anechoic chamber. Filtered “reference” signals are applied to the loudspeaker and the microphone picks up the reflected or “primary” signal. Both the reference and primary data signals are recorded on a Digital Audio Tape and later sampled at 16 kHz and stored to disk for off-line processing. The loudspeaker volume is varied from levels of 75 dB Sound Pressure Level (SPL) to 100 dB SPL, measured at a distance of 0.5 meter. Both the partial adaptive TDNN structure and the NLMS algorithm are applied to the measured data and a number of ERLE curves are obtained for various SPL levels.The algorithm is allowed to converge for 32000 samples and then the average ERLE is obtained from the last 8000 output values.

TIP/TP Performance for the TDNN

The recorded data was applied to the TDNN structure to determine the optimum length for the TDNN section for the highest volume (100 dB SPL) case. The results shown in Figure 6 illustrate that for a system with undermodelling of the impulse length, the TDNN has improved ERLE performance compared to the stand alone NLMS. The best performance comes at approximately 150 taps where the difference between the TDNN and NLMS ERLE value is approximately 5.5 dB.

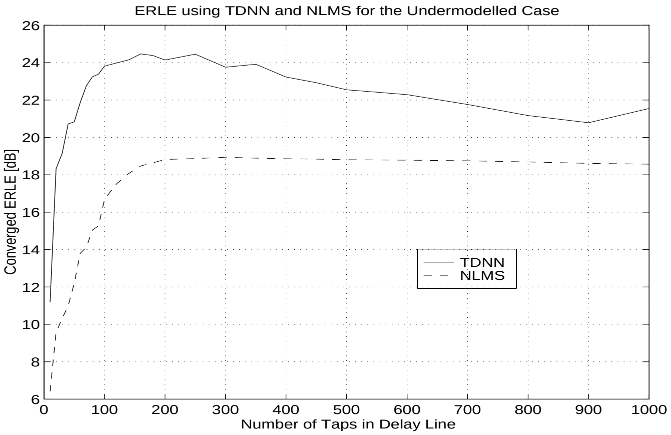

FIGURE 6. Experimental Results. A TDNN is capable of obtaining a better ERLE in an undermodelled state as compared with the NLMS algorithm. Results obtained at a high volume level of 100 dB SPL measured at a distance of 0.5 meter.

Partial Adaptive Structure Performance with Increasing SPL

As shown in Figure 7, the converged ERLE for the partial adaptive structure decreases from a high of 32dB at 80dB SPL to 18.5dB at 100 dB SPL when using the NLMS algorithm. This agrees with results presented in [4] and [8] which show that ERLE is low for low speaker volumes (where acoustic, thermal and DSP related noise are significant) but increases as the reference signal increases, eventually reaching a plateau. Any increase in reference signal level to the loudspeaker after this point results in a *decrease* in the ERLE due to nonlinear distortion. Also shown for comparison is the partial adaptive TDNN algorithm which outperforms the NLMS algorithm at high volume levels. The TDNN section consisted of 150 taps as determined from Figure 6.

.

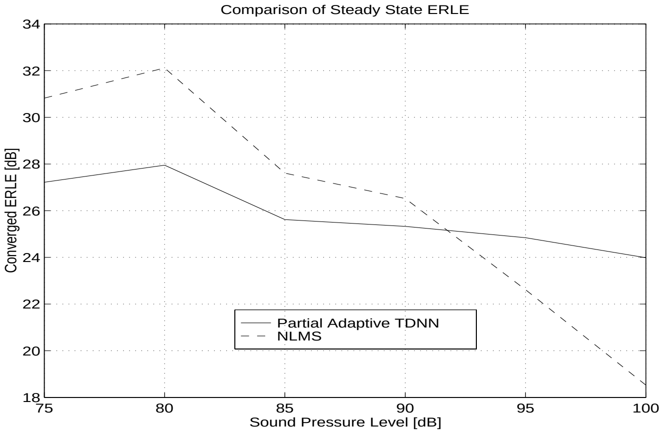

FIGURE 7. Experimental Results. Converged ERLE performance of the partial adaptive TDNN structure compared to the NLMS structure. A 5.5 dB improvement in ERLE can be obtained at high volumes.

The length of the total impulse response is the same for both the partial adaptive TDNN structure and the baseline NLMS structure and is truncated to 600 taps. Note the improvement in ERLE over the NLMS case is significant in the high SPL volume ranges and is greater than 5.5 dB at volume levels in the vicinity of 100 dB SPL.

Convergence

Figure 8 illustrates the ERLE convergence of the partial adaptive TDNN structure compared with the NLMS structure, obtained using data recorded at 100 dB SPL. The convergence rate of the new structure is slightly worse than the NLMS and will affect the tracking performance of the AEC. Methods to reduce this are currently under investigation.

.

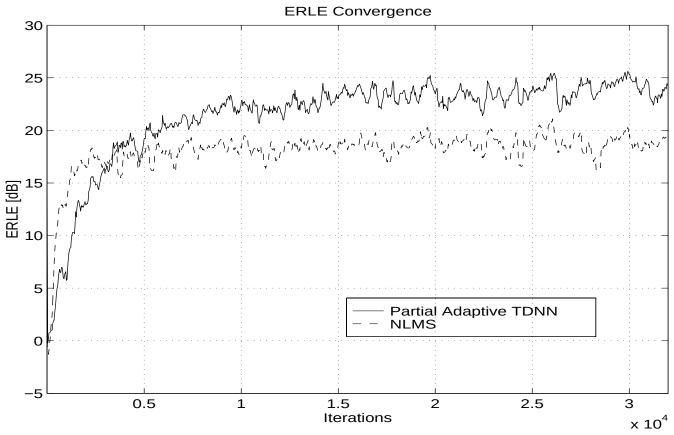

FIGURE 8. Experimental Results. ERLE convergence curves for the partial adaptive TDNN structure and the stand alone NLMS AEC at the highest volume.

CONCLUDING REMARKS

A TDNN structure has been shown to improve the achievable ERLE of a loudspeaker/microphone system at high volumes. This suggests the use of a partial adaptive structure incorporating a short delay TDNN to replace a section of the NLMS filter. The partial adaptive TDNN structure was found to improve the ERLE performance over the NLMS baseline AEC by 5.5dB at high volumes where loudspeaker nonlinearities limit the achievable ERLE. All measurements were performed using real audio components. Although the new structure clearly offers improvements at high volume (i.e. high distortion) levels, it does not quite match the performance of the NLMS structure a low distortion levels. It also has a slightly slower convergence rate, although acceleration techniques were not employed. This is the subject of future research.

REFERENCES

[1] H.F. Olsen, *Acoustical Engineering,* Toronto, D. Van Nostrand Company,Inc., 1964.

[2] X.Y. Gao, W. M. Snelgrove, “Adaptive Nonlinear Recursive State-Space Filters”, *I.E.E.E. Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, Vol. 41, No. 11, Nov. 1994, pp. 760-764.

[3] X. Y. Gao, W. M. Snelgrove, “Adaptive Linearization of a Loudspeaker”, *ICASSP* 1991 Vol. 3, pp 3589-3592.

[4] S. Haykin, *Adaptive Filter Theory*, 2nd ed., Prentice-Hall, Toronto, 1991

[5] M.E. Knappe, R.A. Goubran,”Steady State Performance Limitations of Full-Band Acoustic Echo Cancellers”, *ICASSP* 1994, Adelaide, South Australia, Vol. 2, pp. 73-76.

[6] A.N. Birkett, R. A. Goubran, “Acoustic Echo Cancellation for Hands-free Telephony Using Neural Networks”, *Neural Networks for Signal Processing 1994, IEEE Workshop Proceedings,* Sept. 1994,pp. 249-258.

[7] P. Chang, C. Lin, B. Yeh, “Inverse Filtering of a Loudspeaker and Room Acoustics Using Time-delay Neural Networks”, *Journal of the Acoustic Society of America,* Vol 95, No. 6, June 1994, pp. 3400-3408.

[8] S. E. Kuo, J. Chen, “Multiple Microphone Acoustic Echo Cancellation System with the Partial Adaptive Process”, *Digital Signal Processing* 3, 1993, pp. 54-63.

[9] Y. Pao, *Adaptive Pattern Recognition and Neural Networks,* Addison-Wesley Publishing Company, Inc, 1989.

ACKNOWLEDGEMENTS

The authors wish to thank NSERC, Carleton University, BNR and the Telecommunications Research Institute of Ontario (TRIO) for their financial support.