ACOUSTIC ECHO CANCELLATION FOR HANDS-FREE TELEPHONY USING NEURAL NETWORKS

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: One of the limitations of linear adaptive echo cancellers in hands-free environments is their inability to effectively cancel nonlinearities which are generated mainly in the loudspeaker during large signal peaks. The soft-clipping effect encountered when large signals are applied to the loudspeaker is modelled in a neural network using a piecewise linear/sigmoid activation function. A three layer fully adaptive feedforward network is used to model the room/speakerphone transfer function using the special activation function. This network structure improves the ERLE performance by 10 dB at low to medium loudspeaker volumes compared to a NLMS echo canceller.

INTRODUCTION

A microphone placed next to a loudspeaker in a closed loop provides electro-acoustic feedback which will spontaneously oscillate at some frequency for which the modulus of the gain factor is greater than one. This arrangement exists in all hands-free telephone systems hence adaptive echo cancellation is required to prevent these oscillations while communicating in full-duplex mode.

Limitations of echo cancellers for speakerphones [4],[8] include (a) acoustic, thermal and DSP related noise, (b) inaccurate modelling of the room impulse response (c) slow convergence and dynamic tracking, (d) nonlinearities in the transfer function caused mainly due to the loudspeaker, and (e) resonances and vibration in the plastic enclosure.

To be commercially attractive, convergence times on the order of 100 ms with Echo Return Loss Enhancement (ERLE) on the order of 30 dB are necessary. Fast RLS based adaptive techniques can be used to reduce the convergence time, however, the ERLE is degraded when the input data is severely non-stationary and it has been found [4],[5] that for large filter orders and nonstationary environments, LMS type algorithms will give better overall performance than RLS type algorithms. However, nonlinear techniques must be employed to deal with system nonlinearities and IIR recursive structures must be utilized when poles exist in the room/speakerphone transfer function [6]. In this paper, a tapped delay line feedforward neural network is employed in an attempt to model only the system nonlinearities.

Distortions in the Loudspeaker

A loudspeaker has several sources of nonlinearity including non-uniform magnetic field and nonlinear suspension system [1]. Nonlinear distortion is often a few percent of the output signal and it is desirable to reduce it.A loudspeaker consists of an electrical part and a mechanical part as shown in Figure 1. The electrical part is the voice coil and the mechanical part consists of the cone, the suspension system and the air load. The two parts interact through the magnetic field. The resulting equation of motion [2] is:

$$m\frac{d^2x}{dt} + r_M\frac{dx}{dt} + f_M = Bli \qquad (1)$$

where B is the magnetic flux density in the air gap, l is the length of the voice coil conductor, x is the cone displacement, m is the total mass of the coil, cone and air load and f_M is the force deflection characteristic of the loudspeaker cone suspension system, usually approximated by;

$$f_M = \alpha x + \beta x^2 + \delta x^3 \qquad (2)$$

where α, β and δ are modelling constants and x is the displacement of the voice coil. Suspension system nonlinearity manifests itself as soft clipping at the loudspeaker output and results in odd-order harmonics under large signal conditions.

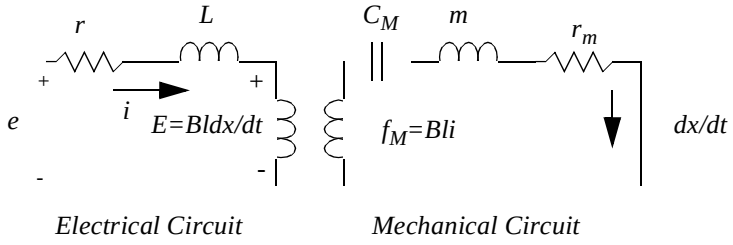


Figure 1. Loudspeaker Electro-mechanical Equivalent Model. e indicates the internal voltage of the generator, r is the total electrical resistance of the generator and voice coil, L is the inductance of the voice coil, i is the amplitude of the current in the voice coil, E is the voltage produced in the electrical circuit by the mechanical circuit. B is the magnetic flux density in the air gap, l is the length of the voice coil conductor, and x is the cone displacement. In the mechanical circuit m is the total mass of the coil, cone and air load. r_M is the total mechanical resistance due to dissipation in the air load and the suspension system. C_M is the compliance of the suspension and f_M is the force generated in the voice coil.

CONVENTIONAL ADAPTIVE ECHO CANCELLER MODELS

Linear Transversal Filter Model

Figure 2a illustrates an acoustic echo canceller (AEC) utilizing a linear adaptive transversal filter to model the room impulse response to cancel the reflected signal. The reflected signal is a combination of room echoes, direct path signals, loudspeaker and microphone transfer functions, and vibration and resonances emanating through the plastics of the speakerphone as illustrated in Figure 2b. The normalized Least Mean Square (NLMS) algorithm [10] is the baseline by which performance of alternative models is measured.

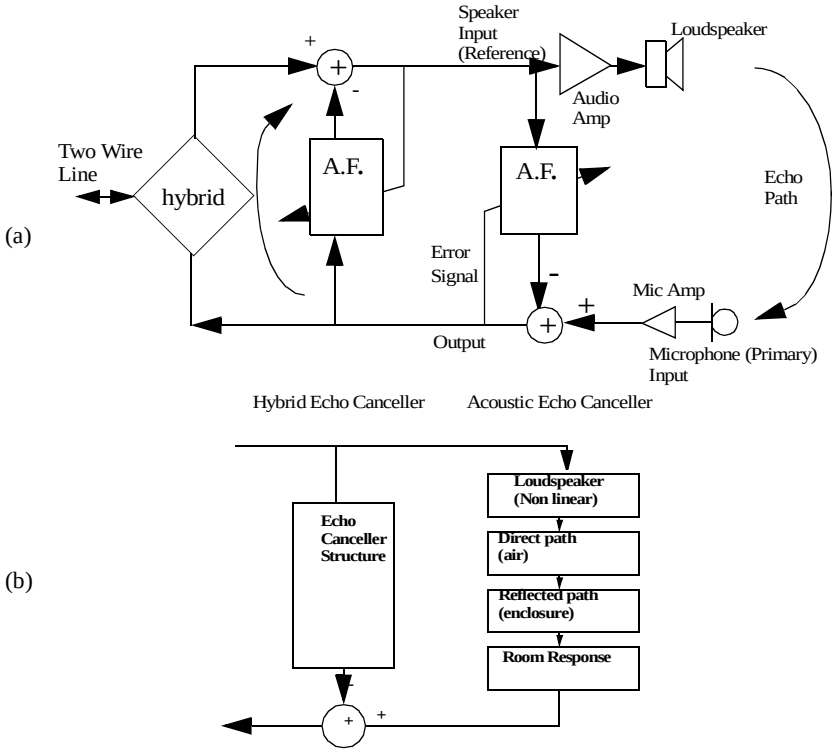


Figure 2. (a) Adaptive Echo Canceller Structure (b) Room/terminal Transfer Function is a combination of Speaker Non-linearities, Direct path, Plastic effects and Room response.

Nonlinear Adaptive Volterra Model

Adaptive volterra filtering can be utilized to deal with loudspeaker nonlinearities [2], however, filter orders greater than 3 are required to effectively model the

speaker transfer function and this very quickly leads to an unmanageably huge model [9]. In fact, during the course of this work, a fully connected 3rd order adaptive Volterra filter with m1=600, m2=600, and m3=50 where m1, m2 and m3 refer to the orders of the linear, quadratic and cubic sections respectively, was constructed in an attempt to model the loudspeaker nonlinearity. The tap updates were based on the LMS algorithm presented in [9] but extended to a cubic system. It was found that no noticeable improvement in converged ERLE could be seen using this technique. Neural networks offer an alternative method of dealing with high order system nonlinearities.

NEURAL NETWORK ECHO CANCELLER MODELS

Three separate adaptive AEC networks were constructed. The first AEC uses a two layer (100,2,1) network placed in series with a 500 tap NLMS adaptive linear filter as shown in Figure 3a. The 100 inputs are obtained from a tapped delay line. The hidden layer neuron has a nonlinear activation function and the output neuron is linear.The neural network in this case is first batch trained on the first 500 points of data obtained at a medium volume and then tested on loud volume data to ensure that the network is not overtrained.

The second AEC uses the same network but in this case, the neural network is placed in parallel with the NLMS adaptive linear filter as shown in Figure 3b.

The third AEC model utilizes a fully adaptive (600,2,2,1) 3 layer feedforward neural network. The 600 inputs are obtained from a tapped delay line. The two hidden layer neurons have piecewise linear/sigmoid nonlinear activation functions and the output neuron is linear. This model is shown in Figure 3c.

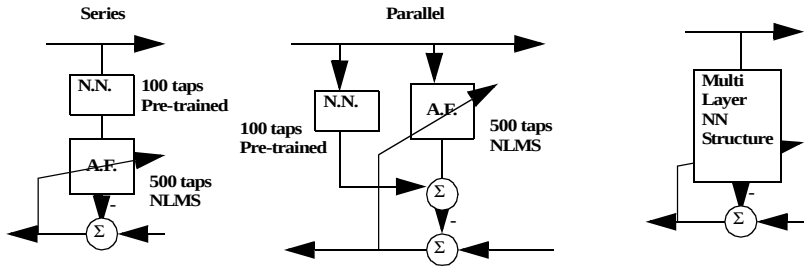
In each neural network, a piecewise linear/tan-sigmoid activation function is used in order to mimic the soft clipping effect and the function response is shown in Figure 4 along with its corresponding delta function. The transfer function is linear below a user definable point and then follows a compressed hyperbolic tangent sigmoid beyond this point such that the output is squashed between +/-1.0. The linear region was set to +/- 0.75 since it was found that this gave good results.

In all cases, the backpropagation algorithm with a normalized step size is employed during the training and tracking phase. The stepsize μ is normalized [10] according to (3).

$$\mu = \frac{\alpha}{\varepsilon + \sum_{i=0}^{M-1} x_i^2} \qquad (3)$$

where α is a number between 0 and 2, and in all cases is set to 0.5. ε is a small positive constant used to prevent the stepsize from becoming too large, *M* is the num-

ber of delay sections in the tapped delay line (i.e. order of the input section) and x_i is the amplitude of the i^{th} delayed element. The stepsize μ is updated after each new sample is shifted into the tapped delay line.



(a) (b) (c)

Figure 3. (a) Series model has a (100,2,1) 100 tap pretrained 2 layer network in series with the NLMS adaptive structure. (b) The parallel model has a (100,2,1) pretrained 2 layer network in parallel with the NLMS structure. (c) The fully adaptive (600,2,2,1) three layer network . In all cases the output neuron is linear and the hidden layers have a piecewise linear / sigmoidal activation function. The inputs are obtained from a tapped delay line.

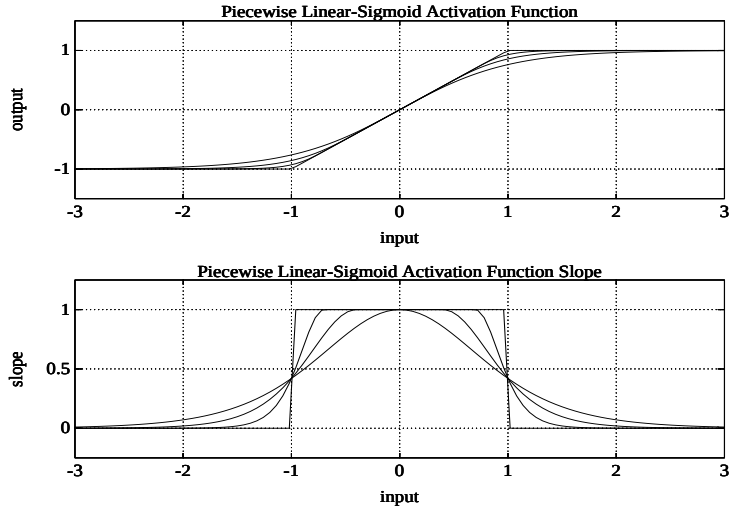


Figure 4. Piecewise linear/sigmoidal activation function and corresponding delta . The linear section with a value of +/-0.75 to +/-0.9 gave the best results in this study.

EXPERIMENTAL SETUP

Figure 5 illustrates the test set-up used to obtain the data. A number of commercially available speakerphones were purchased and modified to allow access to internal signals. The modified speakerphone is placed inside a noise shielded enclosure or 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.

TEST RESULTS

Converged ERLE for NLMS Case

The NLMS algorithm with 600 taps is applied to the measured data and a number of ERLE curves are obtained for various speaker volume levels. The algorithm is allowed to converge for 32000 samples and then the average ERLE is obtained from the last 8000 output values. The results illustrated in Figure 6, show that the converged ERLE is low for low speaker volumes where acoustic, thermal and DSP related noise are significant. This agrees with results presented in [4] and [8]. The ERLE increases as the reference signal increases but reaches a plateau. Any increase in reference signal level to the loudspeaker after this point results in a decrease in achievable ERLE. The NLMS results in Figure 6 are obtained from three different commercially available speakerphones ranging in price form $32 to $120.

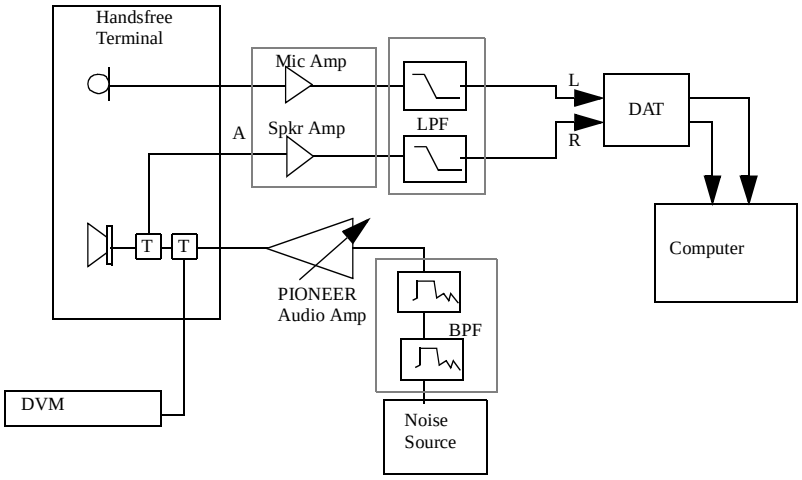


Figure 5. Experimental Setup. Primary and reference signals are recorded on DAT and later sampled to disk.

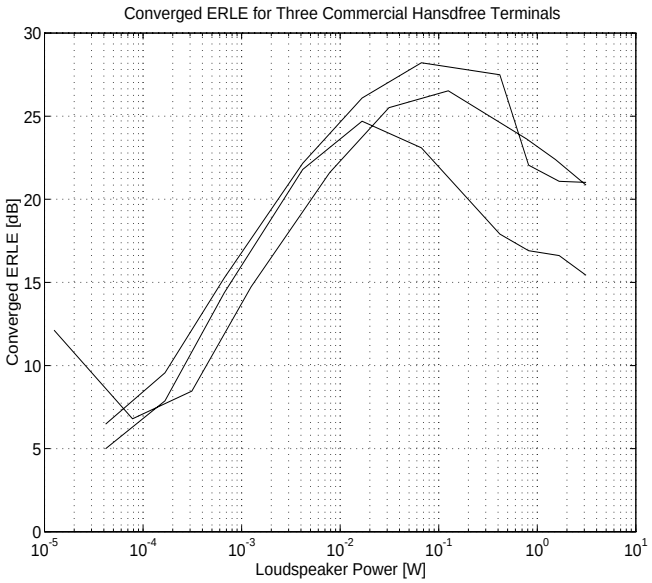


Figure 6. Converged ERLE vs. loudspeaker power for three different commercially available handsfree telephone terminals.. An AEC using the NLMS algoritm shows a decrease in ERLE as the volume increases At low volume levels, noise limits the achievable ERLE.

Convergence Curves for Parallel and Series Models Utilizing Pretrained Neural Networks

The ERLE convergence curves of the series and parallel structures are illustrated in Figure 7. Also illustrated for comparison is the 600 tap NLMS case. The series model has a slightly superior convergence than the NLMS case but eventually settles to the same value of converged ERLE. The parallel structure has a convergence essentially the same as the NLMS case but settles to a lower value of converged ERLE. These results were obtained at a high volume of 0.25 W which is equal to the rated power handling capability of the loudspeaker.

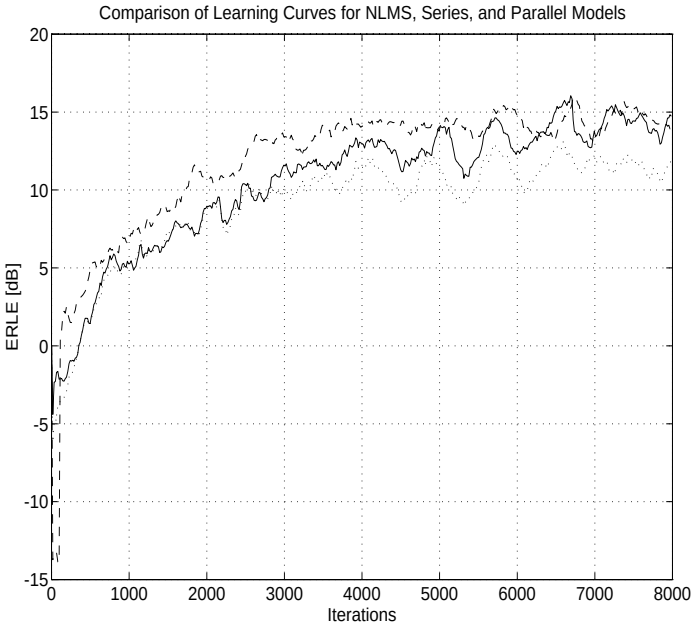


Figure 7. Convergence curves for the Series (dashed line) and Parallel (dotted line) models. The NLMS convergence curve (solid line)is shown for comparison.

Converged ERLE for the Fully Adaptive Three Layer Neural Network

Figure 8 illustrates the performance of the fully adaptive (600,2,2,1) structure compared to the 600 tap NLMS case. The improvement in ERLE over the NLMS case is significant in the low and medium volume ranges and is greater than 10 dB at power levels in the vicinity of 1mW. However, the fully adaptive model does not offer significant improvement at high speaker volumes suggesting that there still exists a deficiency in the modelling of the room/speakerphone transfer function at these volume levels. A total of three speakerphones were tested. Each speakerphone yielded similar results.

.

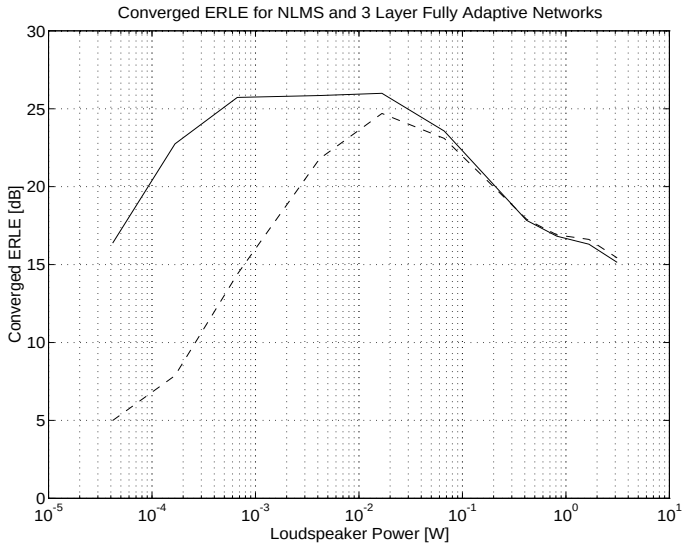


Figure 8. Converged ERLE plot vs. loudspeaker volume using a three layer fully adaptive neural network. Three layer network(solid line) shows over 10 dB improvement in ERLE at low to medium volumes. The NLMS algorithm (dashed line) is shown for comparison.

DISCUSSION OF TEST RESULTS

It has been shown in this paper that a fully adaptive three layer neural network offers significant improvement in converged ERLE in the low to medium volume range where acoustic, thermal and DSP related noise are significant. However, when feedforward structures are utilized at high volume levels, little or no improvement in converged ERLE is observed for filtered noise inputs, and this is confirmed by both the Volterra models and the three neural network models presented in this paper.

It appears that the room/speakerphone transfer function may contain poles when the loudspeaker is at high volumes. This is most likely caused by resonances in the plastics of the speakerphone and to a lesser extent poles in the room transfer function [6]. In order to more accurately model the room/speakerphone transfer function, a recursive structure may be necessary and this is the thrust for future work. NARMAX [11] models, recursive neural networks, and nonlinear state-space filters [12] are all possible candidates.

It is likely that the limitation in converged ERLE at high volumes is a combination of nonlinear effects in the loudspeaker and undermodelling of resonances in the plastic enclosure, and that the limitation due to nonlinearity is being masked by the latter.

SUMMARY

Nonlinear distortions and undermodelling has been found to limit the converged ERLE of acoustic echo cancellation in handsfree terminals. Loudspeaker distortions include nonlinearity in the suspension system which will result in soft clipping at high volumes. A piecewise linear/tan-sigmoid activation function has been developed to more accurately model the soft clipping effect and offers a slight improvement in converged ERLE. A third order Volterra model and three neural network AEC models have been developed which indicate that a purely feedforward tapped delay line structure is not sufficient to accurately model the room/speakerphone transfer function at high volumes resulting in no significant improvement in converged ERLE. However, a 10 dB improvement in converged ERLE can be obtained in the low to medium volume ranges where the primary signal to noise ratio is small. It is proposed that at high volumes resonances in the plastic may be masking the nonlinearity of the speaker and that a recursive structure incorporating poles in the transfer function may be necessary to obtain further improvements in converged ERLE.

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

[3] O. Nerrand, P. Roussel-Ragot, L. Personnaz, G. Dreyfus, "Neural Network Training Schemes for Nonlinear Adaptive Filtering and Modelling", IJCNN 1991 pp I-61 to I-67.

[4] M. E. Knappe, Acoustic Echo Cancellation: Performance and Structures, M. Eng. Thesis, Carleton University, Ottawa, Canada, 1992.

[5] H. Yuan, Dynamic Behavior of Acoustic Echo Cancellation, M. Eng. Thesis, Carleton University, Ottawa, Canada, 1994.

[6] Y. Haneda, S. Makino, Y. Kaneda, "Common Acoustical Pole and Zero Modelling of Room Transfer Functions", I.E.E.E. Transactions on Speech and Audio Proc. Vol. 2, No. 2, April 1994, pp. 320-328.

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

[8] M.E. Knappe, R.A. Goubran,"Steady State Performance Limitations of Full-Band Acoustic Echo Cancellers", Presented at ICASSP 1994, Australia.

[9] C. E. Davila, A. J. Welch, H.G. Rylander, "A Second Order Adaptive Volterra Filter with Rapid Convergence", I.E.E.E. Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-35, No. 9, Sept. 1987, pp. 1259-1263.

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

[11] S. Chen, S. A. Billings, "Representations of Nonlinear Systems: The NARMAX Model", International Journal of Control, Vol. 49, No. 3, 1989, pp. 1013-1032.

[12] D. A. Johns, W. M. Snelgrove, A. S. Sedra, "Adaptive Recursive State-Space Filters Using a Gradient-Based Algorithm", IEEE Transactions on Circuits and Systems, Vol. 37, No. 6, June 1990, pp. 673-684.