COMPARISON OF ACOUSTIC ECHO CANCELLERS BASED ON NEURAL NETWORK

Jin S. Seo and Chang D. Yoo

Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Korea
Tel.: +82-42-869-5470, E-mail: pobi@eeinfo.kaist.ac.kr

ABSTRACT

In order to attenuate nonlinearly distorted echoes in systems that can potentially have feedback from the loudspeaker to the microphone, acoustic echo cancellers based on neural network are considered. This paper compares three acoustic echo cancellers based on different neural network structures: (1) 3-layer TDNN (tapped delay line neural network), (2) PTDNN (parallel-TDNN) and (3) LNTDNN (linear/nonlinear TDNN). The neural network structures performed better than the conventional LMS in terms of the converged ERLE (Echo Return Loss Enhancement), with LNTDNN showing the best result. However, all three neural network structures converged slower than the LMS.

1. INTRODUCTION

In video/tele conference systems, feedbacks from the loudspeaker to the microphone may occur, and these may be heard by the originating talker as a delayed and distorted version of his/her own speech. These feedbacks called echoes are often annoying, and thus these systems will require acoustic echo cancellation (AEC) to prevent feedbacks.

In Figure 1, a typical echo cancellation system is illustrated. The signal received from the far-end is nonlinearly distorted by the loudspeaker, and after bouncing-off various surfaces in the room, it is picked up by the microphone. Before the multipath signal y that is picked up by the microphone is transmitted back to the far-end, the output of the adaptive filter H, denoted by \( \hat{y} \), is subtracted from y. For the sake of simplicity it is assumed that the parameters of the adaptive filter H are updated using the transmitted signal e only during the single-talk period.

In order to suppress the echoes, the parameters of the adaptive filter H must adapt so that the filter H models the echo path. Although the LMS least mean square filter has been widely used because of its simplicity and robustness, it can not sufficiently model the echo path due to the nonlinearity. Thus acoustic echo cancellers based on neural network which can model nonlinearities are considered.

In this paper, we compared the performances of three adaptive structures used in AEC. (1) 3-layer TDNN (tapped delay line neural network), (2) PTDNN (parallel-TDNN) and (3) LNTDNN (linear/nonlinear TDNN). Of the three structures, TDNN has been previously considered in AEC [2, 3] while the other two have not.

2. NONLINEARITIES IN THE LOUDSPEAKER SYSTEM

The differential equation for the mechanical circuit which model a loudspeaker is given by

\[
\frac{d^2 x}{dt^2} + r_M \frac{dx}{dt} + \frac{x}{C_M} = B i
\]

where \( m \) is the total mass of the coil, \( x \) is the cone displacement, \( r_M \) is the total mechanical resistance due to dissipation in the air load and the suspension system, \( C_M \) is the compliance of the suspension, \( B \) is the magnetic flux density in the air gap, \( l \) is the length of the voice coil, \( i \) is the amplitude of the current in the voice coil [4].

The primary source of nonlinear distortions in the loudspeaker comes from nonlinear suspension system. The force deflection characteristic of the loudspeaker cone suspension system can be usually approximated by a third-order polynomial, thus we have

\[
\frac{d^2 x}{dt^2} + r_M \frac{dx}{dt} + \alpha x + \beta x^2 + \gamma x^3 = B i
\]
where $\alpha, \beta, \gamma$ are the coefficients of the polynomial.

The secondary source of nonlinear distortions is non-uniform magnetic field. The flux density is not a constant, instead it is a function of the displacement $x$, which is approximated by a second-order polynomial given by

$$B(x) = B_0 + B_1x + B_2x^2$$

(3)

where $B_0, B_1, B_2$ are the coefficients of the polynomial.

The nonlinear distortions become more important when dealing with low cost (small) loudspeakers. A. N. Birkett states in [3] that nonlinear distortions can be easily 5 to 10% of the total loudspeaker output especially when dealing with small loudspeakers that have low power ratings.

3. NONLINEAR ECHO CANCELLERS

Conventional linear acoustic cancellers, such as the LMS, are inherently limited in attenuating nonlinearly distorted echoes [1]. A way to deal with loudspeaker nonlinearities represented by Equation (2) and (3) is to use adaptive Volterra filtering [4]. Although the Volterra series is both a useful tool for analyzing weakly nonlinear systems and a basis for synthesizing nonlinear filters with desired parameters, its realization is cumbersome and computationally inefficient. Neural network structures which are easier to realize than the Volterra filters are quite capable of modeling nonlinearities [2].

This paper considers three adaptive nonlinear filters based on three different neural networks: (1) 3-layer TDNN (tapped delay line neural network), (2) PTDNN (parallel-TDNN), and (3) LNTDNN (linear/nonlinear TDNN).

In Figure 2, a 3-layer TDNN structure is shown. Henceforth, 3-layer TDNN structure shown in Figure 2 will be called just TDNN. The inputs to this structure are obtained from a tapped-delay line, and at each hidden layer three neurons are used. Tanh-sigmoid is used as the activation function of the neurons at the hidden layers.

In Figure 3, PTDNN (parallel-TDNN) structure [5] is shown. It is composed of $M$ small scale TDNNs that are connected in parallel. In this paper, $M = 2$ case is considered only. The output $\hat{y}$ is the sum of $M$ TDNN outputs, $\{\hat{y}^{(m)}\}_{m=1}^M$. This structure (parallel structure) has an advantage in real-time processing.

![PTDNN (parallel-TDNN) structure](image)

Figure 3: PTDNN (parallel-TDNN) adaptive filter

In Figure 4, LNTDNN (linear/nonlinear TDNN) is shown. This model combines TDNN and the LMS in a parallel form. The LMS and the 2-layer TDNN represent respectively the linear part and the nonlinear part of the structure. Here again, tanh-sigmoid is used as the activation function of the three neurons at the hidden layers.

![LNTDNN (linear/nonlinear TDNN) structure](image)

Figure 4: LNTDNN (linear/nonlinear TDNN) adaptive filter

4. COMPUTER SIMULATION

Computer simulations were performed using white Gaussian noise $N(x, 0, \sigma = \sqrt{10})$ as an input. The signal $x$ was first nonlinearly distorted, then filtered by artificial room impulse response. Nonlinearity was modeled by both quadratic and cubic distortions according to

$$y = \frac{\alpha x + bx^2 + cx^3}{[a] + [b] + [c]}$$

(4)
where $a$, $b$, and $c$ respectively refer to the amplitude of the linear, quadratic and cubic factors [3]. The room reverberation (room impulse response) was modeled by a 200-tap finite impulse response (FIR) filter. In Figure 5, the room impulse response used in the simulation is shown. Since the required number of tapped-delay lines is assumed unknown, 80-tapped delay line was used in each structure. Both the converged ERLE (Echo Return Loss Enhancement) and the convergence rate are used as performance measures. The performance of the LMS [6] was used as a benchmark. The converged ERLE is defined as

$$ERLE(dB) = 10 \log \frac{E[y^2]}{E[e^2]} = 10 \log \frac{\text{var}(y)}{\text{var}(e)} .$$

The ERLE was estimated from 1000 windowed samples.

![Figure 5: Room impulse response](image)

Figure 6 illustrates the behavior of converged ERLE for each structure. The signal-to-distortion ratio (SDR) is calculated as

$$SDR(dB) = 10 \log \frac{\text{var}(x)}{\text{var}(y - x)} .$$

In training the neural network, Levenberg-Marquardt algorithm was used iteratively on 5000 samples. Approximately 1 ~ 2 dB gain in ERLE over that of the LMS was achieved at low SDR. As the SDR increased, the gain in using a neural network structure decreased. Structures TDNN and PTDNN ($M = 2$) performed similarly. Although the complexity of LNTDNN is only slightly higher than that of TDNN, it outperformed the LMS in terms of the converged ERLE whereas the TDNN performed only slightly better. As shown in Figure 6, the converged ERLEs of all structures increased with increasing SDR until SDR=30dB when ERLEs started to saturate. This can be attributed to the fact that the neural network structures do not provide complete model the reverberation (the number of taps in the adaptive filters was less than length of the room impulse response).

Figure 7 shows the convergence behavior of TDNN structure at $SDR = 13.81$ dB. The gradient descent with momentum method was used as the learning algorithm for the TDNN and the gradient descent method was used for the LMS. In order to make a fair comparison, each learning rate was initially adjusted so that the converged ERLEs of both structures were equal. The simulation results showed that the convergence rate of LMS is superior to that of TDNN. Because TDNN requires more time to converge than LMS, it is recommended that it be used in the situations where the room impulse response vary slowly. The convergence behavior of PTDNN structure was found to be similar to that of TDNN.

![Figure 6: Converged ERLE of each structure vs. Signal-to-distortion ratio](image)

![Figure 7: ERLE of TDNN vs. Number of input samples](image)

Figure 8 shows the convergence behavior of LNTDNN structure at $SDR = 13.81$ dB. The gradient descent with momentum method was used as the learning algorithm for the nonlinear part and the gradient descent method was used for the linear part. The learning rate of the LNTDNN was set so that its converged ERLE was equal to that of the LMS. When this was done, LNTDNN converged slower than the LMS but much faster than TDNN.

Figure 9 illustrates the ERLE performance of each structure after training with 5000 samples at $SDR = 13.81$ dB. After training, each structure was tested on 35000 samples. As shown in Figure 9, LNTDNN outperformed all others. In terms of ERLE, the TDNN structure performed slightly better than the LMS. This figure shows that slow convergence rate of neural network can be overcome by the initial training.
5. SUMMARY

By considering the nonlinear distortion in a loudspeaker, we can improve the AEC performance. Three adaptive neural network structures were considered and compared with the conventional LMS. The neural network structures performed better than the conventional LMS in terms of the converged ERLE (Echo Return Loss Enhancement), with LNTDNN showing the best result. However, all three neural network structures converged slower than the LMS. Among the three acoustic echo cancellers based on neural network, LNTDNN performed the best. We conclude that using a hybrid structure (combining LMS and neural network), like LNTDNN, can bring promising results.

Since speech and white Gaussian noise have different characteristics, different results may be obtained using speech. We leave the verification for further work.

6. REFERENCES


