ACOUSTIC ECHO CANCELATION FOR HIGH NOISE ENVIRONMENTS

Amit S. Chhetri[1] *Jack W. Stokes*[2] *Dinei A. Florêncio*[2]

[1]Arizona State University, Tempe AZ 85287, amit.chhetri@asu.edu
[2]Microsoft Research, Redmond, WA 98052, {jstokes,dinei}@microsoft.com

ABSTRACT

Acoustic echo cancellation (AEC) is highly imperative for enhanced communication in noisy environments such as a car or a conference room. In this work, we present a dual-structured AEC architecture that improves both the convergence time and misadjustment of a conventional adaptive subband AEC algorithm in high noise environments. In this architecture, one part performs smooth adaptation while the other part performs fast adaptation; a convergence detector is implemented to facilitate switching between the fast and smooth adaptations. We propose the momentum normalized least mean square (MNLMS) algorithm for smooth adaptation and we implement the NLMS algorithm for fast adaptation. The current architecture provides up to 3-4 dB echo reduction improvement over a conventional adaptive subband AEC algorithm and it helps minimize near-end distortion and artifacts in the post-processed AEC output.

1. INTRODUCTION

Acoustic echo cancellation (AEC) removes the echo captured by a microphone when a sound is simultaneously played through speakers located near the microphone [1]. Many high-noise environments such as noisy conference rooms or lobbies and hands-free telephony in cars require effective echo cancellation for enhanced communication. However, the presence of noise impedes the convergence of the AEC algorithm, which leads to poor echo cancellation. Furthermore, nonlinear post processing techniques such as the use of a center clipper result in noticeable distortion in the near-end speech.

Previous works on AEC in high noise focussed on combined noise and echo reduction ([2] and references therein). One of the approaches in [2] is to preprocess the microphone signal through a noise suppression (NS) algorithm and perform adaptation using the far-end speaker signal that has undergone the same NS operations as the microphone signal. Although, this seems favorable, our experiments revealed that this technique often distorts the echo signal, which hinders the convergence properties of the AEC algorithm. Furthermore, this technique requires perfect synchronization between the microphone and the far-end speaker signals.

In this work, we are concerned with improving the AEC system[1] performance in high noise conditions. We propose a new AEC architecture with the objective of maximizing the echo cancellation of the AEC algorithm. Our motivation is that by maximizing echo cancellation of the AEC algorithm, we can make the post-processing stages milder, and thereby minimize near-end distortion and artifacts. Furthermore, the AEC algorithm makes more use of the signal information than the post AEC algorithms as it takes into account both the phase and the magnitude of the input signals; the post AEC algorithms do not use the phase information of the signals.

A well known property of adaptive filtering algorithms is the trade-off between adaptation time and misadjustment [3]. An effective AEC requires fast adaptation when the echo path changes and smooth adaptation when the echo path is stationary. In this work, we develop a dual-structured AEC architecture where one part of the architecture performs fast adaptation, while the other part performs smooth adaptation. We propose the momentum normalized least mean square (MNLMS) algorithm for smooth adaptation and we perform fast adaptation using the NLMS algorithm. We demonstrate through our experimental results that our proposed architecture provides up to 3-4 dB gain in echo cancellation over the conventional adaptive subband NLMS based AEC algorithm.

This paper is organized as follows. In Section 2, we describe a conventional subband NLMS based AEC algorithm. In Section 3, we describe our proposed AEC architecture. Performance results are discussed in Section 4, and conclusions are provided in Section 5.

2. SUBBAND AEC ALGORITHM

We consider a typical audio-conferencing environment in which the far-end (speaker) signal x played out through the speakers produces an echo at the microphone [3]. In addition to the echo from the speakers, the audio signal y captured by the microphone is also composed of the desired speech s and background noise n. The AEC algorithm cancels the echo from the microphone signal resulting in the output signal e.

In this paper, the conventional adaptive subband AEC algorithm implements a subband NLMS algorithm. The signals are sampled at 16 KHz and they are processed on a frame-by-

[1]The AEC system in this paper implies an AEC algorithm followed by post-processing echo suppression and noise suppression algorithms.

frame basis with each frame measuring 20 ms. To compute the spectrum we use a 320-point modulated complex lapped transform (MCLT) every 20 ms using a 40 ms window. The MCLT is a particular form of cosine modulated filter-bank that allows for perfect reconstruction. The frequency domain spectrum is divided into 320 frequency bins with a bin separation of 25 Hz.

3. PROPOSED AEC ARCHITECTURE

The AEC algorithm performs reasonably well under low noise conditions; as the noise level increases, the adaptation is hindered and the AEC performance deteriorates [3]. In addition, disturbance effects such as time-varying echo paths (due to movements in a room) further reduce the AEC's performance. In such scenarios, one not only requires fast convergence rates, but also low levels of misadjustment.

We propose a dual-structured AEC architecture in which one part performs fast adaptation while the second part performs smooth adaptation. At any given time, a convergence detector is used to decide which of the two parts should be used for the final AEC output signal. In the proposed architecture, we use the NLMS algorithm for fast adaptation and the MNLMS algorithm for smooth adaptation. To explain our architecture, we first describe the MNLMS algorithm.

3.1. Momentum Normalized Least Mean Square Algorithm

The MNLMS algorithm proposed in this paper is a variant of the momentum LMS (MLMS) algorithm, which, was first used in digital communication for high speed adaptive equalization [4]; the MLMS algorithm was shown to provide faster and smoother convergence than the LMS algorithm [5]. The MNLMS algorithm corresponds to a *second-order* adaptive algorithm in that two previous weight vectors are combined at each iteration to obtain the updated weight vector [5]:

$$\begin{aligned}\mathbf{W}(f,k+1) &= \mathbf{W}(f,k) + 2\mu\frac{\mathbf{X}(f,k)E^H(f,k)}{(\mathbf{E}\left[\|X(f,k)\|^2\right]+\delta)} \\ &\quad +\alpha\left[\mathbf{W}(f,k)-\mathbf{W}(f,k-1)\right], \quad (1)\end{aligned}$$

where f is the frequency index, k is the frame index, $\mu > 0$ is the adaptation step-size, $-1 < \alpha < 1$ is the momentum factor, $\mathbf{X}(f,k) = [X(f,k), \cdots, X(f,k-L+1)]^T$ is the far-end speaker signal vector with $X(f,k)$ denoting the far-end speaker signal for subband f and frame index k, L denotes the regression model order, $\mathbf{W}(f,k) = [W_1(f,k), \cdots, W_L(f,k-L+1)]^T$ denotes the weight vector, $\mathbf{E}$ denotes an expectation, and $\delta > 0$ is a regularization term. Also, $E(f,k)$ is the AEC output: $E(f,k) = Y(f,k) - \mathbf{W}^H(f,k)\mathbf{X}(f,k)$, where $Y(f,k)$ is the microphone signal and H denotes the Hermitian operation. The third term in the summation of (1) is called the momentum term, since by adding a fraction of the weight increment of the previous time-step, we provide momentum to the adaptive process. Note that for $\alpha = 0$, the MNLMS algorithm reduces to the NLMS algorithm.

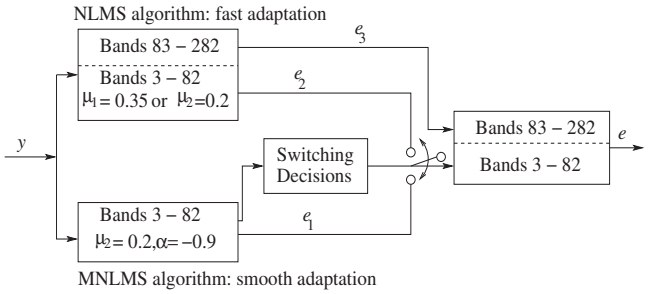

Fig. 1. Proposed dual-structured AEC architecture.

It was noted in our experiments that under high noise conditions and $\alpha > 0$, MNLMS performs poorly in comparison with the NLMS algorithm. This is because the weight update equation is largely disturbed by high background noise; any error made during the previous weight update step is propagated to the future time-steps due to the momentum term. Interestingly, when $\alpha < 0$, the MNLMS algorithm performs better in terms of misadjustment than the NLMS algorithm (when using the same step-size for both algorithms). This is because by using a negative α, the update in weights of the previous time-step are rendered unreliable (due to high noise) unless there is a strong feedback in the future time-step for this weight update. This builds a smoothing effect in the MNLMS algorithm which makes it more resilient to noisy conditions than the NLMS algorithm. This motivates us to use the MNLMS algorithm for the smoothing part in our proposed AEC architecture, which is described next.

3.2. Dual-Structured AEC Architecture

We propose a dual-structured AEC architecture as shown in Fig. 1. In this architecture, two streams of AEC algorithms operate in parallel. The upper stream implements the NLMS algorithm for fast adaptation while the lower stream implements the MNLMS algorithm for smooth adaptation; a convergence detector is implemented to switch between the two streams. The advantage of this architecture is that we can switch between fast and smooth adaptation depending on room conditions. Note that both streams operate independently, i.e., there is no exchange of information between the two streams.

It was found through experiments that the improvement in AEC performance of the lower stream over the upper stream was dominated by adaptation in the frequency bins $3-82$ only. Thus, in our architecture, we operate the lower stream over only the frequency bins $3-82$. This helps reduce the computational cost as we do not have to implement a full-band AEC algorithm in the lower stream.

At each frame k, the AEC output signal e_1 of the lower stream is processed by a convergence detector (described in Section 3.3) to determine if the echo canceler of the lower stream has converged. If convergence is detected, we use e_1 for the bins $3-82$ of the final AEC output signal e, otherwise the bins $3-82$ of the AEC output signal of the upper stream

(i.e., e_2) are used. The frequency bins 83-282 of e always correspond to e_3 i.e., to the frequency bins 83-282 of the AEC output signal of the upper stream.

To further improve the performance of the AEC system, we implement the AEC algorithm of the upper stream with two step-sizes, i.e. μ_1=0.35 and μ_2=0.2, that were chosen through rigorous experimentation with real data. We use μ_1 for fast adaptation and μ_2 when smooth adaptation is required but the MNLMS algorithm has not converged. The decision to switch between μ_1 and μ_2 is made using a separate convergence detector that is built into the AEC algorithm of the first stream. The AEC algorithm of the lower stream implements an MNLMS algorithm with a step-size of μ_2.

To summarize, we begin the adaptation process with the AEC algorithm of the upper and lower streams operating with step-sizes of μ_1 and μ_2, respectively. Initially, the AEC algorithms of both the streams will be in the learning phase (i.e. not converged). However, as the upper stream converges faster than the lower stream, we combine e_2 and e_3 to obtain e during the initial phase. Furthermore, upon convergence of the AEC algorithm in the upper stream, μ_1 is reduced to μ_2 to perform smooth adaptation. When the AEC algorithm in the lower stream converges, e_1 and e_3 are combined to obtain e. Finally, whenever a change in the echo path is detected, we switch to the faster adaptation stream and continue using it until the AEC algorithm in the lower stream reconverges.

3.3. Convergence Detector

An important component of our proposed architecture is to be able to switch between fast and smooth adaptation depending on the convergence conditions of the AEC algorithm. To achieve this, we use the orthogonality property of adaptive algorithms: when the echo canceler has converged, the AEC output signal must be orthogonal to the speaker signal [3]. This property was used to develop a double talk detector in [6]. We adopt the double talk detection algorithm of [6], but use it only as a convergence detector. Further, instead of operating the convergence detector in the time domain, we operate it in the subband domain; this is explained next.

The cross correlation between the AEC output $E_1(f,k)$ of the lower stream at frame k and the speaker signal $X(f,k-i)$ at frame $k-i$ ($i=0,\cdots,L-1$) for frequency bin f is defined as

$$\rho^i(f,k) = \frac{P^i_{XE_1}(f,k)}{P^i_X(f,k)P_{E_1}(f,k)}, \qquad (2)$$

where $P_{E_1}(f,k)$, $P^i_X(f,k)$, and $P^i_{XE_1}(f,k)$ are updated using an exponential weighting recursive algorithm [6]:

$$\begin{aligned} P^2_{E_1}(f,k) &= \lambda P^2_{E_1}(f,k-1) + (1-\lambda)|E_1(f,k)|^2 \\ |P^i_X(f,k)|^2 &= \lambda |P^i_X(f,k-1)|^2 + (1-\lambda)|X(f,k-i)|^2 \\ P^i_{XE_1}(f,k) &= \lambda P_{XE_1}(f,k-1) + (1-\lambda)\cdot \\ &\qquad X(f,k-i)E^H_1(f,k). \end{aligned} \qquad (3)$$

Here, λ is an exponential weighting factor generally set as $0.95 < \lambda \leq 1$ for slowly time varying signals. Using (2), we define the average cross correlation (ACC) as $\bar{\rho}(f,k) \triangleq \frac{1}{L}\sum_{i=0}^{L-1}\rho^i(f,k)$. For reliable convergence decisions, the ACC is computed only for the frequency bins 13-82 (325 Hz - 2.05 KHz) where speech signal is dominantly present.

At each frame k, we compare $\bar{\rho}(f,k)$ to a threshold ρ_{Th} for f=13,11,$\cdots$,82. If the inequality $\bar{\rho}(f,k) \leq \rho_{Th}$ is met for more than half of the total frequency bins considered (i.e., 70/2=35), we declare that the AEC has converged, otherwise we declare that either the AEC has not converged or the echo path has changed. The convergence threshold is typically set to be slightly larger than $\bar{\rho}(f,k)$ in its steady state [6].

4. EXPERIMENTS AND RESULTS

We tested the performance of our proposed AEC architecture on real data collected from a small office-room (10x8x10ft). The data was recorded at 16 KHz sampling rate. To evaluate the AEC performance quantitatively, we consider only the single talk case; the double talk case was evaluated through listening tests. We analyzed the algorithm's performance quantitatively on the basis of echo return loss enhancement (ERLE) in dB, which is given as, $\text{ERLE}(k) = 10\log_{10}\left[\frac{\mathbf{E}\{y^2(k)\}}{\mathbf{E}\{e^2(k)\}}\right]$. For the single talk case, the far end speaker signal was first recorded under low noise conditions; at 16.5 s ($k = 825$) movements were introduced in the room to cause a change in the echo path. Office background noise was collected and then synthetically added to the far-end signal to produce the microphone signal at an echo-to-noise ratio of 7 dB. After processing the microphone signal through the AEC, the background noise was subtracted from the AEC output signal and the result compared with the noise-free microphone signal. This was done to evaluate the true performance of the proposed architecture unhindered by noise. Thus, the terms $y(k)$ and $e(k)$ in the ERLE formulation correspond to the noise-free microphone and AEC output signals, respectively.

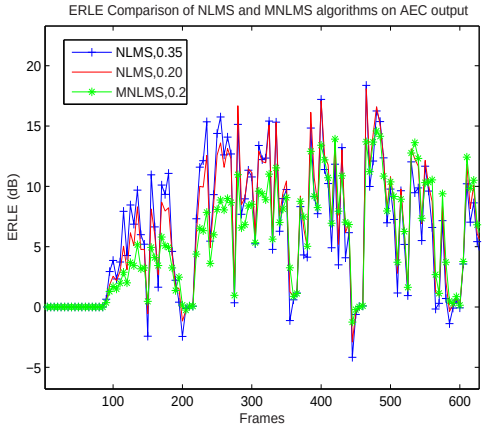

Fig. 2. ERLE comparisons for NLMS and MNLMS algorithms on the AEC output.

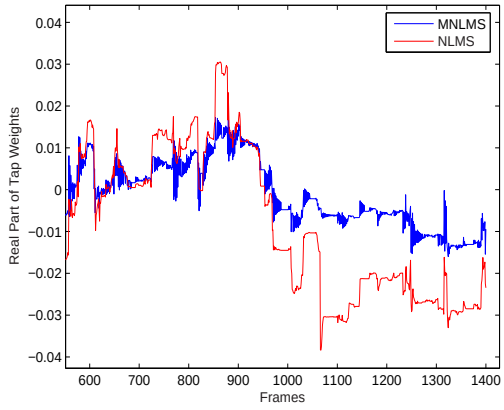

Fig. 3. Comparison of the real part of the first tap weight for the NLMS and MNLMS algorithms.

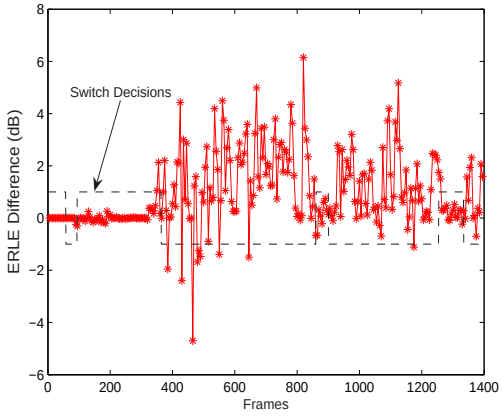

Fig. 4. ERLE difference between the the proposed architecture and the conventional subband adaptive AEC algorithm on the AEC output in room noise.

In this work, we use $\mu_1 = 0.35$, $\mu_2 = 0.2$, and $\alpha = -0.9$. Fig. 2 depicts the ERLE curves for the NLMS (with μ_1 and μ_2) and the MNLMS (using μ_2 and α) algorithms. It can be seen that the adaptation is fastest with NLMS,0.35, followed by NLMS,0.2, and MNLMS,0.2. However, as we approach $k = 500$, the ERLE is largest for MNLMS,0.2, followed by NLMS,0.2, and NLMS,0.35.

Fig. 3 compares the real part of the first tap weight for the NLMS and MNLMS algorithms to demonstrate the difference between the two adaptive algorithms. It can be seen that the tap weight for the NLMS algorithm fluctuates more rapidly than that for the MNLMS algorithm. This suggests that the MNLMS algorithm is more resilient to noisy conditions than the NLMS algorithm. The difference in evolution of the tap weights for the MNLMS and NLMS algorithms is the primary reason to operate the AEC algorithms in the upper and lower streams independently for frequency bins $3 - 82$

We compare the results of our proposed AEC architecture with that of a conventional adaptive subband AEC algorithm described in Section 2 using a step-size of 0.35. Fig. 4 shows the ERLE difference on the AEC output between our proposed architecture and the conventional AEC algorithm. The dashed line indicates the switching decisions along time; a value of 1 indicates that the upper stream is processed while a value of -1 indicates that the lower stream is processed.

Initially, the AEC algorithms in both streams are converging; however, as the upper stream converges faster, we use its output as the final AEC output (during frame indices 1-363). Furthermore, as the conventional AEC algorithm also implements a step-size of 0.35, its output is equivalent to the output of our parallel AEC architecture. As a result, we do not see an ERLE gain between our AEC architecture and the conventional AEC algorithm. At about $k = 321$, the AEC algorithm of the upper stream converges; consequently, the step size of the AEC is changed to 0.2 for smoother convergence. This results in an ERLE gain between the parallel architecture and the conventional AEC algorithm during the frame indices $321 - 363$. At $k = 364$, the lower stream also converges. At this stage, we switch to the lower stream i.e., we use e_1 for frequency bands $3-82$ of e. This leads to an ERLE gain of up to 4 dB over the conventional AEC algorithm. At around $k = 825$, there is a movement in the room, which is detected by the convergence detector. At this time, we shift to the upper stream. Eventually, when the AEC in the lower stream reconverges, we shift back to the lower stream. Thus, the parallel architecture helps to obtain both fast adaptation and low misadjustment, which results in the improved performance of our proposed architecture over the conventional subband adaptive AEC algorithm.

5. CONCLUSIONS

A new dual-structured architecture for the AEC system is proposed to improve both the convergence time and misadjustment of a conventional AEC algorithm in high noise environments. The architecture provides up to 3-4 dB improvement in echo reduction over the conventional subband adaptive AEC algorithm for a small computational overhead.

6. REFERENCES

[1] C. Breining et. al., "Acoustic echo control: An application of very-high-order filters," *IEEE Signal Processing Magazine*, vol. 16, pp. 42–69, July 1999.

[2] R. L. B. Jeannés, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: A survey," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 808–820, Nov. 2001.

[3] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 4 ed., 2001.

[4] J. G. Proakis, "Channel identification for high speed digital communications," *IEEE Trans. Automat. Control*, vol. 19, pp. 916–922, Dec. 1974.

[5] S. Roy and J. J. Shynk, "Analysis of the momentum LMS algorithm," in *IEEE Trans. ASSP*, pp. 2088–2098, Dec. 1990.

[6] H. Ye and B. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Trans. Communications*, vol. 39, pp. 1542–1545, Nov. 1991.