REGRESSION-BASED RESIDUAL ACOUSTIC ECHO SUPPRESSION

Amit S. Chhetri[1] *Arun C. Surendran*[2] *Jack W. Stokes*[2] *John C. Platt*[2]

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

ABSTRACT

In this paper, we propose a novel regression-based algorithm for suppressing the residual echo present in the output of an acoustic echo canceller (AEC). We learn a functional relationship between the magnitudes of many frames of the speaker signal and the magnitude of the echo residual, per subband. We estimate and track the parameters of this function using adaptive algorithms (e.g. NLMS). We show that this approach can be interpreted as a rank-1 approximation to a more general regression model, and can address shortcomings of the earlier approaches based on correlation analysis. Preliminary results using linear regression on magnitudes of real audio signals in both mono and stereo situations demonstrate an average of 7 dB of echo suppression over the AEC output signal under a wide variety of conditions without near-end signal distortion. The framework is general enough to promise even further reductions in the future.

1. INTRODUCTION

The echo reduction provided by an acoustic echo canceller (AEC) is often inadequate for most applications. This insufficiency is caused due to computational constraints that force the adoption of filter lengths in AEC that are much shorter that the room response. Various methods have been employed to suppress the residual echo. Simple techniques such as coring have been traditionally used, albeit with significant near-end speech distortion. Most techniques try to estimate the power spectral density (PSD) of the residual echo, and remove this using Weiner filtering [1, 2] or spectral subtraction [3].

Some methods estimate PSD based on long-term reverberation models of the room [3] require some knowledge of the room configuration and/or impulse response. Other more recent works estimate residual echo PSD using either through the “mismatch transfer function” (the part of the room response that has not been estimated by AEC) [2], or through “coherence analysis” [1]. These methods, either directly or indirectly, depend on correlation analysis - accurate estimation of the cross-correlation between the speaker signal and the residual signal. In a subband system, only the DFTs of the windowed signals are available, so the cross-correlations can only be calculated approximately [1]. While coherence analysis based on a single block of data can lead to biased estimates, using multiple blocks of data can alleviate this problem, albeit leading to frequent overestimation in the presense of near-end speech [1]. Further, the multi-block approach assumes that the frames of the speaker signal are uncorrelated, which is almost never true.

In this paper, we propose regression-based echo suppression (RES) to reduce the echoes in the residual of an AEC signal. Instead of relying on correlation analysis, we propose to directly estimate (per sub-band) the magnitude (or energy) of the short-term spectrum of the AEC residual signal in terms of the magnitudes (or energies) of the short-term spectra of the speaker signal using parameterized relations. In general, regression models can easily capture complex empirical relationships while providing flexibility. We employ these models for their flexibility, rather than their physical accuracy, and hence do not run into some of the problems that occur in correlation-based systems. These models do not need prior knowledge about room reverberation. RES is described fully in Section 2. In Section 3 we show that the model parameters can be easily estimated and tracked using adaptive algorithms. We also show how the RES algorithm can be easily extended to stereo residual echo suppression (Section 3.1). RES provides over 7dB of echo suppression beyond that of AEC under a variety of real conditions (we will present results in Section 4). Thus, our proposed model is both simple to implement and very effective.

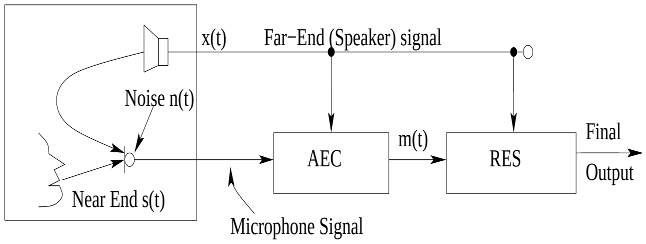

Figure 1: Block diagram showing the role of the Regression-based residual acoustic Echo Suppression (RES) system.

2. REGRESSION MODEL

In RES, we wish to directly estimate the amount of residual echo in each frame of AEC output. We achieve this by modeling the empirical relationship between the speaker signal and the echo residual. The output of the AEC $m(t)$ can be expressed as

$$m(t) = x(t) * h_l(t) + s(t) + n(t) \tag{1}$$

where $s(t)$ is the desired near-end signal at the microphone, $x(t)$ is the far-end or speaker signal, $n(t)$ is the ambient noise, and $h_l(t)$ is the uncompensated part of the room impulse response (see Fig. 1). The echo residual after AEC, $r(t)$, is

$$r(t) = x(t) * h_l(t), \tag{2}$$

where $*$ denotes convolution. In the frequency domain, this is expressed as:

$$R(f) = X(f)H_l(f). \tag{3}$$

This expression holds true only when we consider infinite duration signals. In reality, the signals are processed on a frame-by-frame basis (typically of 20 ms duration) and the true relationship between the short-term frames are complex. In general, the current frame of the residual signal can be expressed in terms of the current and past speaker signal frames:

$$R(f,t) = g_\Theta\left(X(f,t), X(f,t-1), \cdots, X(f,t-L+1)\right) \quad (4)$$

where f and t represent the frequency and time index respectively, g represents an unknown function, Θ is the set of parameters of the model, and L depicts the model order. Once a good estimate of $R(f,t)$ is obtained, it can be subtracted from the residual echo.

2.1. Magnitude Based Regression Model

Typically, a room impulse response lasts a few hundred milliseconds. Depending on the number of taps, the AEC is able to model and cancel the effect of the relatively early echoes. The AEC residual can reasonably be assumed to contain a part of the early echo and most of the late-echoes, also called long-term room response, or late reverberation. The late reverberation consists of densly packed echoes that can be modeled as white noise with an exponentially decaying envelope [4]. This, combined with the belief that the AEC captures a significant part of the phase information, leads us to believe that whatever phase information is left behind will be very difficult to track. Instead, we propose that the magnitude of the short-term spectrum of the echo residual be expressed in terms of the magnitudes of the current and previous frames of the speaker signal. In this paper we use a linear model, although more complex models can be used:

$$|R(f,t)| \approx \sum_{i=1}^{L} w_i |X(f,t-i)| \quad (5)$$

where w_i are the regression coefficients for the magnitude model. Squaring both sides of (5) gives

$$\begin{aligned} |R(f,t)|^2 &\approx \left(\sum_{i=1}^{L} w_i |X(f,t-i)|\right)^2 \quad (6) \\ &= \sum_{i=1}^{L}\sum_{j=1}^{L} w_i w_j |X(f,t-i)||X(f,t-j)| \end{aligned}$$

Notice that Eq. (6) includes cross power terms which take into consideration the inherent correlation in the speaker signal. Thus, it is more powerful than either the power regression model

$$|R(f,t)|^2 \approx \sum_{i=1}^{L} \nu_i |X(f,t-i)|^2, \quad (7)$$

or PSD methods that assume independence between frames [1]. Our experiments show that the magnitude regression model performs better than the power regression model.

2.1.1. A Rank-1 Interpretation

In fact, a more general regression model is

$$|R(f,t)|^2 \approx \sum_{i=1}^{L}\sum_{j=1}^{L} \beta_{ij} |X(f,t-i)||X(f,t-j)|, \quad (8)$$

where each $w_i w_j$ in Eq. (6) has been replaced with β_{ij}. Even though the model in Eq. (8) is more general than that in Eq. (5), it does not perform as well in both echo reduction and convergence rate. There are two reasons for this. The general model in Eq. (8) involves L^2 independent parameters β_{ij} compared to L independent parameters w_i in Eq. (6). Further, Ω (where $\Omega(i,j) = \beta_{i,j}$) may not be full-rank i.e., it has a rank k that is between 1 and L. This may cause the estimates to be too noisy and hence affect both its performance and convergence. Looking at Equations (6) and (8), we can see that a rank-1 approximation of Ω can be:

$$\Omega \approx \mathbf{w}\mathbf{w}^T \quad (9)$$

where $\mathbf{w} = [w_1 w_2 \cdots w_L]^T$. An interesting question is: how can we expand the rank-1 model of Ω up to its true rank? A rank-2 approximation of Ω can be $\Omega \approx \mathbf{w}\mathbf{w}^T + \mathbf{h}\mathbf{h}^T$. Under ideal circumstances, an extension of this model may lead to an eigen decomposition of Ω with $\mathbf{w}\mathbf{w}^T$ representing its largest eigen component, $\mathbf{h}\mathbf{h}^T$ representing the next largest, and so on. Since applying RES once leads to estimating $\mathbf{w}$, it is possible that a second application of RES on the remaining signal may estimate $\mathbf{h}$, and each successive application of RES estimates the next eigen component of Ω, and so on. We hope to perform a more rigorous analysis of this formulation in the future. For now, based on the cursory analysis, we can intuitively presume that repeated application of RES, up to its full rank will lead to successive reduction in echo residual. This is borne out empirically from our experiments, with a second RES step supplying an echo reduction of about 2-5 dB beyond a first RES step.

3. ADAPTIVE RES ALGORITHM

The regression coefficients w_i are a function of the room environment and change as the room environment changes. There is evidence that the time-frequency envelope of the long-term reverberation does not depend on the source-receiver locations in a given room [3, 4]. Hence, there is reason to believe that the longer term echo behavior of a room is less sensitive to small source-receiver location changes compared to early echoes (which the AEC tries to model).

In any event, RES must adaptively update the coefficients. In our RES algorithm, we use a magnitude regression-based NLMS adaptive algorithm[1] [7].

In the following, we describe the details of the RES algorithm; a summary of the algorithm is provided in Table 1. In that table, $\mathbf{w}(t)$ is a weight vector at time t, $X(f,t)$ is the subband speaker signal at subband frequency f and time t, $M(f,t)$ is the subband AEC residual signal at subband frequency f and time t,and $P(f,t)$ is the estimated speaker signal power at subband frequency f and time t.

We set $\mathbf{w}(0)$ and $X(f,t)$ for $t \leq 0$ to be zero vectors. In order to improve convergence, we initialize $P(f,t)$ with the energy in the first frame of the far-end signal. For each time t, we predict the residual signal using the speaker signal and the weights; the predicted residual signal magnitude $\hat{R}(f,t)$ is then subtracted from the magnitude of the current frame of the AEC residual $|M(f,t)|$ to suppress the residual signal component in the microphone signal. The error signal $E(f,t)$ is used to update the weights for time $t+1$. In the gradient calculation, we smooth P using a first order IIR model; α is a smoothing constant which is typically set to a small value $0.05 \sim 0.1$.

The final RES output is given as

[1] RLS or Kalman filters can be also be used.

Initialize:

$\mathbf{w}(0) = \mathbf{0}$

$X(f,t) = 0$ for $t \leq 0$

$P(f,0) = \|X(f,1)\|^2$

For each frame t = $1, \cdots, \infty$*:*

Predict residual signal magnitude

$\hat{R}(f,t) = \sum_{i=1}^{L} w_i(t)|X(f,t-i)|$

Compute the error signal

$E(f,t) = \max(|M(f,t)| - \hat{R}(f,t), N_F(f,t))$

Compute the smoothed far-end signal power

$P(f,t) = \alpha P(f,t-1) + (1-\alpha)\|X(f,t)\|^2$

Compute normalized gradient

$\nabla(t) = \frac{-2E(f,t)|X(f,t)|}{P(f,t)}$

Update the weights

$\mathbf{w}(t+1) = \mathbf{w}(t) - \frac{\mu}{2}\nabla(t)$

Table 1: Residual Echo Suppression Algorithm.

$$B(f,t) = E(f,t)\exp(j\phi) \tag{10}$$

where $\phi = \angle M(f,t)$ is the phase of the AEC output signal. The adaptation is performed only in the absence of near-end speech. We typically choose a small step size so that the residual signal estimate $\hat{R}(f,t)$ is mostly smaller than $|M(f,t)|$. In case $\hat{R}(f,t)$ exceeds $|M(f,t)|$, we multiply the step size μ by a small factor λ, where typically $1 < \lambda < 1.5$. This is to ensure the positivity of $E(f,t)$ as much as possible. Whenever the difference between $|M(f,t)|$ and $\hat{R}(f,t)$ becomes lower than the noise floor, we set $E(f,t)$ to the noise floor. This is depicted in Figure 2. This helps in reducing any artifacts such as musical noise in the RES output. The noise floor is calculated using the minimum statistics noise estimation technique provided in [8].

The regression order, L, is chosen according to the room size. Since higher frequency signal components are absorbed better than lower frequency signal components [4], we use a relatively smaller value of L at higher frequencies. In this work, we choose $L = 10, 13$, and 16 for sub-bands 1-72 (lower frequencies) and $L = 6, 8$ and 10 for sub-bands 73-280 (higher frequencies), for small, medium, and large rooms respectively. Each sub-band spans 25 Hz.

The RES algorithm is similar in operation to AEC, but only acts on the magnitudes of the signal. So the complexity of RES is one quarter that of AEC. The computational cost is further reduced by using fewer number of taps for lower frequencies.

3.1. Application to Stereo AEC

The RES algorithm can be extended to stereo AEC in two ways, both involving two passes of the regression. In the first approach, the model can be applied to the AEC output based on the left speaker signal in the first pass, and then the the right speaker signal in the second pass. Alternatively, the first pass can use the sum of the two speaker signals and the second pass can use the difference.

Stereo AEC has problems with correlations between the channels: RES naturally handles these correlations by removing them in two passes.

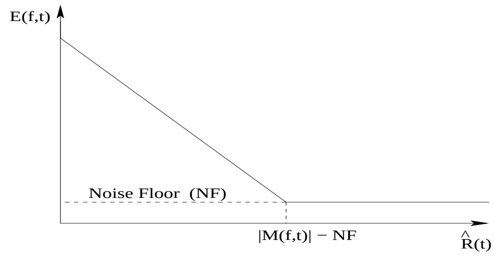

Figure 2: Post processing of the error signal. The error signal is set to the noise floor if it falls below the noise floor. This helps in reducing any artifacts in the RES output.

4. RESULTS

4.1. Database: actual recordings under various scenarios

We tested the performance of RES on real data collected under various scenarios. The data was recorded at 16Khz sampling rate. To compute the spectrum we used a 320-point modulated complex lapped transform (MCLTs) [6] every 10ms using a 20ms window. MCLT is a particular form of cosine modulated filter-bank that allows for perfect reconstruction. FFTs can easily be used instead of MCLTs without changing any other procedure in this paper. We used two different rooms - a medium-sized conference room (14x9x10ft) and a small office-room (10x8x10ft) - with either speech or music playing over the speakers at different interference levels. The AEC is performed using a NLMS algorithm operating on complex subband values. The number of taps are chosen as described in Section 3. The presence of near-end speech is detected by a double-talk detector. After the AEC, we ran a rank-2 RES (i.e. two runs of the algorithm) using the linear regression on magnitudes (described in Section 2.1). We analyze the RES performance on the basis of echo return loss enhancement (ERLE) in dB, which is given as, $ERLE(t) = 10\log_{10}\left[\frac{E\{m^2(t)\}}{E\{r^2(t)\}}\right]$. The ERLE is calculated in the absence of the near-end signal (all frames with a double talk score of < 0.2 were automatically chosen). On the average, we achieved more than 7dB of echo residual reduction (4.2dB after rank-1 RES). With power regression the average residual reduction was 5.5dB. In some cases we achieved more than 12dB suppression.

Audio evaluation of the RES output signals indicated minimal distortion in the near-end signal. The reader is invited to verify this by listening to sample results from our experiments at **http://research.microsoft.com/users/acsuren/aes.aspx**.

Now we will present some specific examples to illustrate the performance of RES.

Figure 3 shows speech signal recorded in the medium-sized room with speech signal playing over the speakers. The AEC output is shown in the upper plot of Figure 3. The average signal-to-residual energy after AEC is about 4dB. Two major regions which are dominated by the echo are marked with the words “echo” in the upper plot (not all segments with echoes are marked). The signal after RES is shown in the lower plot. We can clearly see a significant reduction in the echo in these regions - overall a 10dB reduction is achieved. Figure 4 shows how the magnitudes are tracked for a segment of the residual. The data shown is for the 25th (from lowest) subband.

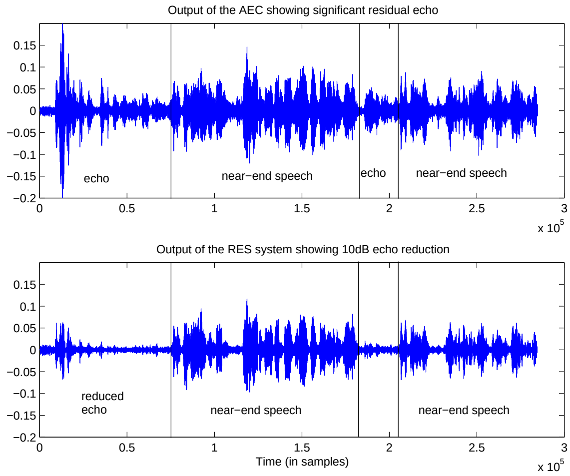

Figure 3: Signals before and after RES showing more than 10dB residual echo reduction.

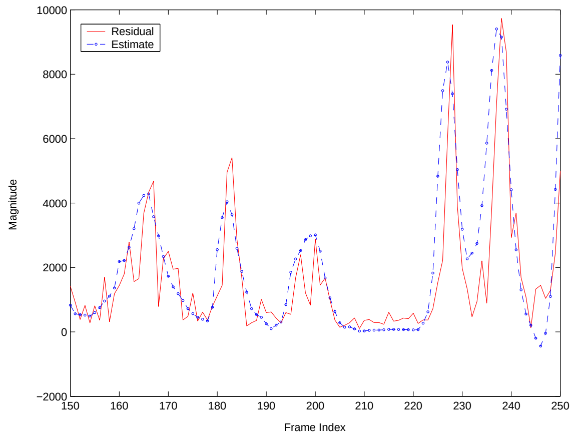

Figure 4: Example showing RES tracking residual echo magnitudes when near-end speech is not present.

The next example is recorded in a small conference room where the far-end signal is music. Figure 5 depicts the ERLE gains over the AEC output that were obtained with the power (dotted line) and magnitude (solid line) regression. It is clear that the magnitude regression significantly outperforms power regression in this case (7.18dB vs 3.65dB). In most cases, magnitude regression is as good as, or better than, power regression model.

Finally, we present preliminary results on a stereo echo-suppression task. The AEC output is once again processed by two runs of RES - the first regression was using the left speaker signal, and then the next was done using the right speaker signal. We obtained an echo suppression of 5 dB and 8 dB with the first and second RES processing, beyond the AEC.

5. SUMMARY

We have presented a simple, yet effective residual acoustic echo suppression system based on regression between the echo residual and the far-end signal. In this paper we present a specific version of the system using linear regression on spectral magnitudes, which automatically takes into consideration the correlation between the frames of the far-end signal. The parameters of the model are tracked easily using adaptive algorithms. Multiple applications of RES can be used to estimate a more general model. This model is easily extendible to stereo-RES. Results on various real signals show that the system provides over 7dB of sustained echo suppression on the average beyond that of AEC with minimal artifacts and/or near-end speech distortion.

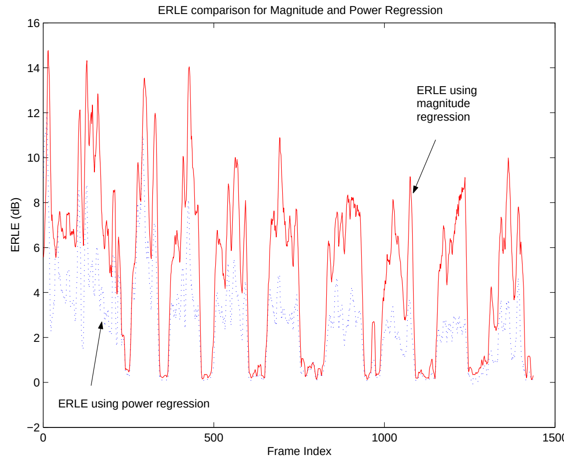

Figure 5: Comparing ERLE gains (in dB) per frame for magnitude (solid line) and power regression (dotted line). The data was recorded in a small room with speech as a desired near-end signal and music as the far-end signal.

6. REFERENCES

[1] G. Enzner, R. Martin and P. Vary, “Unbiased residual echo power estimation for hands free telephony”, ICASSP ’02, pp. 1893-1896, Orlando, Florida, May 2002.

[2] M. Kallinger and K. Kammeyer, “Residual echo estimation with the help of minimum statistics”, IEEE Benelux Signal Processing Symposium, Leuven, Belgium, March 2002.

[3] K. Lebart, *et. al.*, “A New Method Based on Spectral Subtraction for the Suppression of Late Reverberation from Speech Signals”, Audio Engineering Society Issue 4764, 1998.

[4] J-M. Jot, *et. al.*, “Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model”, Audio Eng. Soc. 103rd Convention, New York, 1997.

[5] C. Avendano, *et. al.* , “STFT-Based Multi-channel Acoustic Interference Suppressor”, ICASSP ’01, Utah, May 2001.

[6] H. Malvar, “A modulated complex lapped transform and its applications to audio processing”, ICASSP ’99, pp. 1421-1424, Orlando, Florida, May 2001.

[7] S. Haykin, “Adaptive Filter Theory”, Prentice Hall, 4th Edition, September 2001.

[8] R. Martin, ”Spectral subtraction based on minimum statistics,” Proc. EUSIPCO-94, pp. 1182-1185, Edinburgh, 1994.