NONLINEAR RESIDUAL ACOUSTIC ECHO SUPPRESSION FOR HIGH LEVELS OF HARMONIC DISTORTION

Diego A. Bendersky

University of Buenos Aires
Buenos Aires, Argentina

Jack W. Stokes, Henrique S. Malvar

Microsoft Research
Redmond, WA 98052, USA

ABSTRACT

Linear adaptive filters are often used for Acoustic Echo Cancellation (AEC) but sometimes fail to perform well in notebook computers and inexpensive telephony devices. Low-quality speakers and poorly-designed enclosures that produce vibrations often generate harmonic distortion, and this nonlinear effect degrades the performance of linear AEC algorithms considerably. In this work, we present a new AEC architecture that consists of a linear, subband adaptive AEC filter followed a nonlinear residual echo suppression (RES) stage specifically designed to address harmonic distortion. In addition to suppressing the residual echo in the primary subband, the proposed model also suppresses the residual echo in a window of bands surrounding the higher order harmonics. Results show considerable improvement over other proposed algorithms, and the new algorithm has much lower implementation costs compared to nonlinear AEC models based on Volterra filters and a previously proposed, nonlinear residual echo suppression algorithm.

Index Terms— acoustic echo cancellation, echo suppression, nonlinear distortion, nonlinear filters, nonlinear acoustics

1. INTRODUCTION

An Acoustic Echo Cancellation (AEC) system is a critical component in every full-duplex, speech communication system. Its purpose is to remove the echo captured by the microphone when a signal is played through the speakers without degrading the near-end speech. Although linear adaptive filters have often proved to be adequate solutions to the AEC problem in high quality hardware, they do not perform as well in common, inexpensive laptop computers or telephony devices which introduce nonlinear distortion in the echo. Common sources of nonlinear distortion include low-quality speakers, overpowered amplifiers and poorly-designed enclosures; even modest nonlinear distortion can degrade the performance of linear AEC models considerably. Applications such as hands-free telephony and videoconferencing are particularly problematic due to high loudspeaker volume levels. In laptop computers we have found that high loudspeaker levels often lead to a nonlinear effect known as Harmonic Distortion (HD). Under this effect, signals with high power on particular frequencies produce an increase in the power of frequencies that are exact multiples of the fundamental band. Some laptop computers have lightweight, loose enclosures and high-power speaker signals produce vibrations and reverberances of the entire case generating harmonic distortion.

Several algorithms have been proposed for addressing these problems including nonlinear AEC models and linear Residual Echo Suppression (RES) models. The first group include Volterra filters [1, 2], power filters [3], saturation curve-based predistorters [4] and neural networks [5]. Although these models have been successful in some cases, the large amount of variables and the high order of the operations involved lead to long convergence times and are very expensive in terms of computing requirements. In practical scenarios, RES methods are often preferred. These RES algorithms take the output of the AEC as input and try to predict and further suppress the residual echo. Some of these algorithms include center clipping [6] and linear RES algorithms [7, 8, 9]. In general these methods are more aggressive in the sense that they can reduce the echo further at the expense of some near-end voice distortion in double-talk situations.

Recently, Kuech and Kellerman proposed [10] a nonlinear RES algorithm using a frequency-domain power filter model of the acoustic echo path. Alternatively, in this work we propose another nonlinear, frequency-domain RES algorithm that specifically addresses the problem of Harmonic Distortion in speech signals by modeling inter-frequency dependencies of the speaker and residual echo signals. Compared to the power filter-based, nonlinear RES model, the proposed Harmonic Distortion RES (HDRES) algorithm is very efficient in terms of computational costs since only a single transform of the speaker signal is computed.

This paper is organized as follows. In section 2, the system architecture is presented, and the proposed HDRES algorithm is described in section 3. Numerical results comparing HDRES to several other previously proposed architectures are provided in section 4, and conclusions follow in section 5.

2. SYSTEM ARCHITECTURE

The proposed system architecture is shown in figure 1 and consists of a linear, subband AEC algorithm followed by the nonlinear, HDRES algorithm. The general idea behind this architecture is that we should allow the AEC algorithm to cancel as much echo as possible with the magnitude and the *phase* of the reference signal before disregarding the phase in the HDRES stage. In this paper, we use the modulated complex lapped transform (MCLT) [11], a particular form of a cosine modulated filter-bank that allows for perfect reconstruction, to transform the time domain signals to the frequency domain but any frequency domain transform (e.g. STFT) can be used. The MCLT also allows low-delay architectures when combined with encoders based on the lapped orthogonal transform such as G.722.1 [11]. The AEC is a frequency-domain linear adaptive algorithm that performs per-band time prediction, and the estimated echo $\hat{D}(\kappa, m)$ can be described as:

$$\hat{D}(\kappa, m) = \sum_{t=0}^{T-1} W_L(t, m) X(\kappa - t, m) \qquad (1)$$

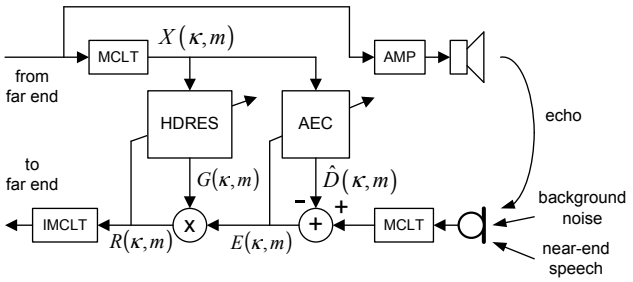

Fig. 1. Diagram of the proposed architecture.

where W_L is a complex weight matrix for the linear AEC, X is the complex transform of the speaker signal, κ is the frame index, m is the frequency band, and T is the number of taps considered. The HDRES filter is a magnitude-only predictor that addresses the HD effect and will be described in the following section.

3. HARMONIC DISTORTION RESIDUAL ECHO SUPPRESSION ALGORITHM

The HDRES problem can be modeled as a noise suppression problem: if we consider the residual echo as noise, an additive signal plus noise model can be used, where the near-end speech plus background noise is the signal and the residual echo is the noise. The input to the HDRES algorithm, $E(\kappa, m)$, which is also the output of the linear AEC, is

$$E(\kappa, m) = D_r(\kappa, m) + S(\kappa, m) + N(\kappa, m) \tag{2}$$

where $D_r(\kappa, m)$ is the true, residual echo signal, $S(\kappa, m)$ the near-end signal and $N(\kappa, m)$ the background noise. Under this assumption, we further suppress the residual echo per band using a magnitude regression model based on the residual fundamental band and the harmonic frequencies as:

$$R(\kappa, m) = G(\kappa, m)E(\kappa, m). \tag{3}$$

The real valued gain, $G(\kappa, m)$, is given by:

$$G(\kappa, m) = \frac{\max\{\bar{E}(\kappa, m) - \beta\bar{D}_r(\kappa, m), \bar{N}(\kappa, m)\}}{\bar{E}(\kappa, m)} \tag{4}$$

with smoothed magnitudes estimates of the AEC output, $\bar{E}(\kappa, m)$, residual echo, $\bar{D}_r(\kappa, m)$, and noise floor, $\bar{N}(\kappa, m)$, computed using recursive averages as:

$$\bar{E}(\kappa, m) = (1 - \alpha)\bar{E}(\kappa - 1, m) + \alpha|E(\kappa, m)| \tag{5}$$

$$\bar{D}_r(\kappa, m) = (1 - \alpha)\bar{D}_r(\kappa - 1, m) + \alpha|\hat{D}_r(\kappa, m)| \tag{6}$$

$$\bar{N}(\kappa, m) = (1 - \alpha)\bar{N}(\kappa - 1, m) + \alpha|\hat{N}(\kappa, m)|. \tag{7}$$

In addition, $|\hat{N}(\kappa, m)|$ is the estimate of the magnitude of the noise floor for frame κ and subband m computed by minimum statistics [14], β can be used to tune the "aggressiveness" of the algorithm [10], α controls the amount of smoothing, $\hat{D}_r(\kappa, m)$ is the estimated, residual echo, and $R(\kappa, m)$ is the complex output of the HDRES. It should be noted that multiplying by the real valued gain $G(\kappa, m)$ affects only the magnitude of each subband, but not the phase, and (3) can be viewed as a spectral subtraction [13] approach similar to [8] when $\alpha = 1$. The magnitude regression model can be used since the residual phase information is difficult to predict and is noncritical for speech intelligibility [12]. Given the microphone signal contains background noise, spectral subtraction based on microphone signal estimation also suppresses the background noise introducing unpleasant musical noise. To reduce the modulations of the background noise, we apply a spectral flooring to the gain computation in (4) [14]. Following the algorithm previously proposed for linear RES [8], we compute $G(\kappa, m)$ in the results in section 4 based on the instantaneous magnitudes of $|E(\kappa, m)|$, $|\hat{D}_r(\kappa, m)|$, and $|\hat{N}(\kappa, m)|$ with $\alpha = 1$ in (5), (6), and (7), respectively. Finally, although we propose using the magnitude-based regression estimate of the harmonic distortion, the HD problem can also be addressed from a minimum mean square error perspective by replacing the magnitudes in (5), (6), (7) (e.g. $|E(\kappa, m)|, |\hat{D}_r(\kappa, m)|, |\hat{N}(\kappa, m)|$) with the square of the corresponding magnitude values [10] and using power regression instead of magnitude regression below. In the magnitude regression model, we need to compute $|\hat{D}_r(\kappa, m)|$ which is discussed next.

3.1. Residual Echo Estimation Model

Neglecting the delay effect of the acoustic echo and considering correlated speaker and residual echo signals under harmonic distortion, the speaker signal at frequency f affects the residual echo signal at frequencies f, $2f$, $3f$, etc. To describe this effect, we propose a linear additive model:

$$|\hat{D}_r(\kappa, m)| = \sum_{i=1}^{M}\sum_{j=1}^{H}\sum_{k=-K}^{K} \delta(i, j, k, m)W_R(i, j, k)|X'(\kappa, i)| \tag{8}$$

where

$$\delta(i, j, k, m) = \begin{cases} 1 & \text{if } i \times j + k = m \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

and i is the fundamental frequency band, M is the number of subbands, j is the harmonic, H is the number of harmonics considered, $2K + 1$ is the length of the harmonic search window with index k, $W_R(i, j, k)$ are the parameters of the HDRES model and $X'(\kappa, i)$ is a transformed version of the speaker signal at frame index κ, frequency i. When dealing with a discretized version of the signal, the frequency domain transform of each harmonic can span several bands and can be displaced with respect to the integer division/multiplication; typically we set $K = 1$ to handle echo leakage from adjacent subbands. It should be noted that searching for the potential bands for each harmonic (i.e. $\delta(i, j, k, m) = 1$) can be implemented very efficiently by considering a fundamental frequency then calculating the window of bands surrounding each possible harmonic. In other words, the actual implementation of (8) is sparse.

In the proposed algorithm, the regression is performed using the frequency-domain transforms of a single frame of the speaker signal and the microphone signal. Ideally, the magnitude regression in (8) would be with respect to time (i.e. multiple speaker frames [8]) in addition to the harmonics, but this is prohibitive in terms of CPU consumption. Furthermore, the speaker and the residual echo signals must be correlated, but the combination of the acoustic echo path and the hardware produces a delay between both signals which may be difficult to estimate in personal computers since operating systems are not hard real-time. An approximation which addresses both problems is to compute the regression using a normalized transformation based on the delayed speaker signal, $|X'(\kappa, m)|$, weighted

by the linear AEC taps weights as:

$$|X'(\kappa,m)| = \sum_{t=0}^{T-1} L(t,m)|X(\kappa-t,m)| \quad (10)$$

where the corresponding, normalized weighting factor $L(t,m)$ is:

$$L(t,m) = \frac{|W_L(t,m)|}{\sum_{j=0}^{T-1}|W_L(j,m)|}, \quad (11)$$

and W_L is the weight matrix of the linear AEC algorithm. This transformation leads to better results for the HDRES algorithm compared with a fixed scalar delay (for example, taking the maximum over all the weights of the AEC).

Vibration of the enclosure is usually only produced with high powered signals; hence, harmonic distortion is only noticeable when the magnitude of some frequencies is high. To avoid incorrect adaptation of the model when HD effect is not present, we introduce an adaptive threshold for the speaker signal power in order to predict if a given frequency would produce harmonics. This threshold is based on the average power of the speaker signal. As the main goal of the algorithm is to attenuate high-powered frequencies, we also apply a threshold to the microphone and the residual signal: that is, we apply the filtering process when both the speaker and the microphone signals for the particular band are above the given thresholds. Also, we adapt the weights of the model only when the residual signal is not negligible.

3.2. Model Adaptation

Since the HDRES algorithm uses a linear model with respect to harmonics in the transformed speaker signal, we can use any linear adaptive algorithm to update them; in this case we use the normalized, least mean square (NLMS) algorithm [15] as:

$$\begin{aligned} \xi(\kappa,m) &= |E(\kappa,m)| - |\hat{D}_r(\kappa,m)| \\ W_R(i,j,k) &\leftarrow W_R(i,j,k) + \frac{\mu}{\bar{P}(\kappa,m)}|X'(\kappa,m)|\xi(\kappa,m) \end{aligned}$$

where $m = ij + k$, μ is the step size, and the average power in the transformed speaker signal is:

$$\bar{P}(\kappa+1,m) = (1-\rho)\bar{P}(\kappa,m) + \rho|X'(\kappa,m)|^2.$$

As with other RES methods, the algorithm is sensitive to double talk detection: if adaptation occurs when near-end voice is present, even for a short period of time, the near-end voice distortion increases considerably. Since most double talk detectors are based on averaged signal statistics (e.g. speaker, microphone, error), it takes a few frames in order to detect a change. Besides, spurious short single-talk segments can be incorrectly detected in the middle of long double-talk segments. To cope with these effects, we propose two simple, yet effective mechanisms: adaptation rollback and hysteresis control. Considering that the last adaptation steps before a change from single-talk to double-talk were incorrect, adaptation rollback consists of discarding the last T_1 adaptation steps before the double-talk detector transitions from single-talk to double-talk. This mechanism is implemented by keeping a window of the last T_1 instances of the weight matrix. On the other hand, hysteresis control is simply implemented by turning off the double-talk detector (i.e., going from double-talk to single talk and enabling adaptation) only when T_2 consecutive frames are classified as single-talk.

Denomination	AEC	RES	ERLE (dB)
AEC_{Freq}	Freq. Lin.		11.75
AEC_{Time}	Time Lin.		13.94
AEC_{Pow}	Time Power		14.21
AEC_{Freq}+MRRES	Freq. Lin.	Time Reg.	14.98
AEC_{Freq}+HDRES	Freq. Lin.	Harm. Dist.	20.28
Max (SNR)			30.49

Table 1. ERLE comparison for different AEC/RES methods. Max represents the Signal-to-Noise Ratio.

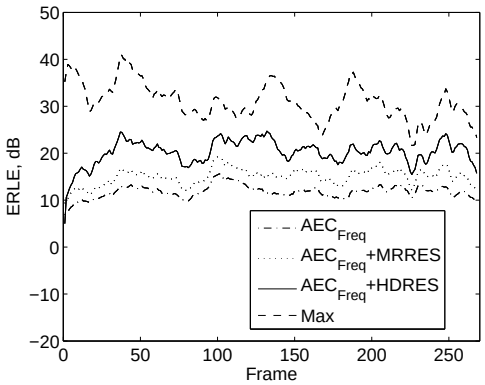

Fig. 2. Evolution of ERLE over time.

4. EXPERIMENTAL RESULTS

To test our algorithm, we carried out several tests and we compare the results with other known algorithms in terms of echo suppression and other performance measurements. Table 1 compares echo attenuation in using ERLE (echo return loss enhancement) for several AEC architectures. The 16 kHz sampled data was recorded in a noisy environment, using a notebook with a high level of harmonic distortion. Figure 2 shows the evolution of the ERLE over time. AEC_{Freq} is the frequency-domain NLMS-based algorithm described in [11]. MRRES is a frequency-domain, linear RES algorithm based on magnitude regression and described in [8], AEC_{Time} is a time domain AEC implementation where the taps are updated using NLMS, and AEC_{Pow} is an implementation of a time-domain power filter with NLMS adaptation [3]. Finally, HDRES is the RES algorithm presented in this paper. Max is the Signal-to-Noise Ratio (SNR) of the microphone signal and provides an upper bound for the ERLE. The AEC_{Freq} + HDRES combination outperforms the other methods with an improvement of 8.53 dB over AEC_{Freq} alone. Figure 3 shows the time-domain output signals obtained with AEC_{Freq} and AEC_{Freq}+HDRES in subplots a) and b), respectively.

The parameter H indicates the potential number of harmonic frequencies that each frequency may affect. Figure 4 shows the ERLE obtained for different values of H. Here, it is possible to see that the ERLE values grow asymptotically, and a value of 10 is adequate for most of the tests, with a maximum difference of more than 2 dB compared to a value of 2. Figure 5 shows a spectrogram representing a segment of conversation with single-talk and double-talk intervals; bars indicate single-talk detection. We can see that with the introduced enhancements, false positives (i.e. double-talk

detected as single-talk) are reduced considerably at the expense of more false negatives that do not affect the performance of the algorithm.

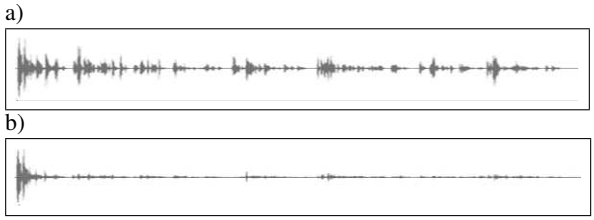

Fig. 3. Time-domain residual echo plots for a) AEC_{Freq} and b) AEC_{Freq}+HDRES.

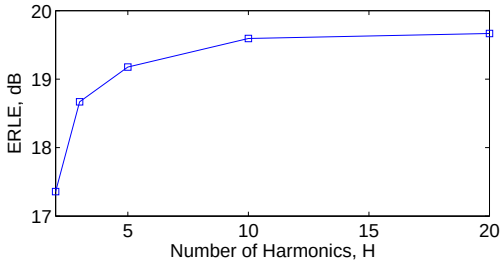

Fig. 4. ERLE as a function of the number of harmonics considered

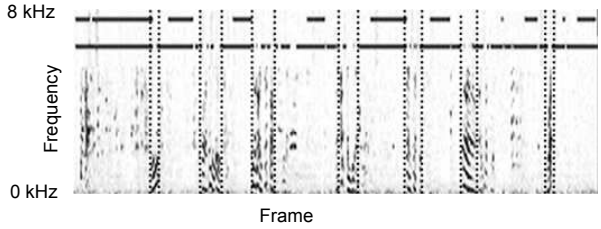

Fig. 5. Spectrogram of the AEC+HDRES algorithm output. Segments between dashed lines represent near-end speech. The bars indicate single-talk segments with (upper) and without (lower) double-talk detector enhancements.

5. CONCLUSIONS

Residual acoustic echo suppression algorithms offer a good compromise between linear AEC models and nonlinear AEC models. Due to the significant number of parameters in nonlinear AEC models, in particular those using Volterra filters, reduced convergence speed and high computational complexity may prohibit these solutions for many scenarios. RES algorithms do not suffer from convergence problems since the linear AEC quickly converges to a first order solution in a timely manner. Furthermore, previously proposed linear RES algorithms have shown the ability to reduce the effects of system nonlinearities due to the speakers and the enclosure. The *nonlinear*, harmonic distortion RES algorithm proposed in this paper as well as the nonlinear RES algorithm based on power filters presented in [10] can achieve better echo suppression with more sophisticated models. The power filter RES algorithm [10] first computes L frequency domain transforms of the speaker signal raised to the l^{th} power, (i.e. $x^l(k)$) for $l = 1 : L$. Next, an orthogonal representation of these L transforms is then computed. For the HDRES algorithm we propose, a single transform of the speaker signal is computed resulting in significant reduction of computational resources. Thus, the HDRES algorithm falls in the middle of the complexity range of echo cancellation/suppresion architectures spanning: linear AEC, linear AEC plus linear RES, linear AEC plus HDRES, linear AEC plus nonlinear power filter-based RES, and nonlinear AEC. As a result, it can be a realistic solution for many of today's teleconferencing products. Preliminary experiments show an improvement of perceptual quality without degrading the near-end speech, but additional formal listening tests are needed to quantify such improvement.

6. REFERENCES

[1] A. Guérin, G. Faucon, and R. Le Bouquin-Jeannès, "Nonlinear acoustic echo cancellation based on volterra filters," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 672–683, 2003.

[2] S. Im, "A normalized block lms algorithm for frequency-domain volterra filters," in *Proc. SPW-HOS '97*, Washington, DC, USA, 1997, p. 152, IEEE Computer Society.

[3] F. Kuech, A. Mitnacht, and W. Kellerman, "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," in *Proc. ICASSP '05*, 2005, pp. 105–108.

[4] H. Dai and W. Zhu, "Compensation of loudspeaker nonlinearity in acoustic echo cancellation using raised-cosine function," *IEEE Trans. on Circuits and Systems. II: Express Briefs*, vol. 53, no. 11, pp. 1190–1194, 2006.

[5] A.B. Rabaa and R. Tourki, "Acoustic echo cancellation based on a recurrent neural network and a fast affine projection algorithm," in *Proc. IECON '98.*, 1998, pp. 1754–1757.

[6] O. M. M. Mitchell and D. A. Berkley, "Reduction of long-time reverberation by a center-clipping process," *Journal of the Acoustical Society of America*, vol. 47, no. 1, pp. 84, 1970.

[7] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21–32, 1998.

[8] A.S. Chhetri, A.C. Surendran, J.W. Stokes, and J.C. Platt, "Regression-based residual acoustic echo suppression," in *Proc. IWAENC '05*, Eindhoven, The Netherlands, 2005.

[9] O. Hoshuyama and A. Sugiyama, "An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo," in *Proc. ICASSP '06*, 2006, pp. 21–32.

[10] F. Kuech and W. Kellerman, "Nonlinear residual echo suppression using a power filter model of the acoustic echo path," in *Proc. ICASSP '07*, 2007, pp. I–73–I–76.

[11] H.S. Malvar, "A modulated complex lapped transform and its applications to audio processing," in *Proc. ICASSP '99*, 1999, pp. 1421–1424.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[13] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[14] R. Martin, "Spectral subtraction based on mininum statistics," in *Proc. EUSIPCO '94*, 1994, pp. 1182–1185.

[15] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 2001.