ACOUSTIC ECHO CONTROL BASED ON TEMPORAL FLUCTUATIONS OF SHORT-TIME SPECTRA

Alexis Favrot, Christof Faller

Illusonic LLC
Lausanne, Switzerland

Markus Kallinger, Fabian Küch, Markus Schmidt

Fraunhofer IIS
Am Wolfsmantel 33
91058 Erlangen, Germany

ABSTRACT

Recently echo suppressors operating in a short-time spectral domain were introduced without a need for an adaptive FIR filter for echo path estimation. The echo path is modeled with merely a delay and a real-valued gain at each frequency, representing direct sound and early reflections. A weakness of this approach is, that these gains are biased when there is noise in the microphone signal. This paper proposes an improved method to estimate these gains, based on temporal fluctuations of short-time spectra. It is shown that the proposed technique results in estimated gains without a bias when there is noise in the microphone signal.

***Index Terms*—** Acoustic Echo Cancellation, Echo Path Estimation, Echo Suppression, Adaptive Filter.

1. INTRODUCTION

Acoustic echoes in a tele-communication system arise whenever far-end sound from the loudspeaker is picked up by the microphone in the same room as illustrated in Figure 1. The far-end signal x, emitted by the loudspeaker, travels to the microphone both directly and through reflected paths. Thus, the microphone signal y does not only comprise the local near-end speech and noise w but also the echo which is thus fed back to the user on the far-end,

$$y[n] = h[n] * x[n] + w[n], \tag{1}$$

where h is the room impulse response and $*$ denotes convolution.

Issues caused by acoustic echoes are potential instability (howling) and annoyance of hearing one's own echo. Echoes are the more annoying the more they are delayed [1]. While analogue telephony has low delay, modern digital based telephony or tele-conferencing systems often have a round trip delay of a few dozen milliseconds.

To overcome the issues caused by acoustic echoes, echo control is needed for hands free telecommunication systems. Ideally, an acoustic echo canceler (AEC) [2] enables full duplex communication since it only removes the echoes from the microphone signal while other signal components remain unchanged. AECs use adaptive FIR filters of lengths of up to several hundred milliseconds to model the acoustic echo path, which results in high computational complexity. Usually AECs are combined with a non-linear post-processor [3] (subband suppressor) to eliminate residual echoes which occur whenever the echo path changes or when there are non-linearities.

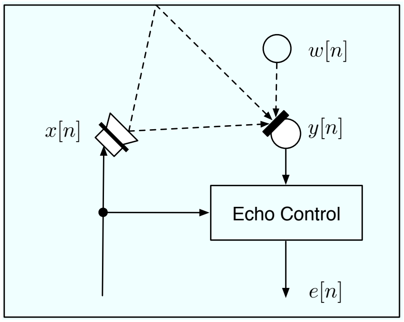

Fig. 1. A general setup illustrating acoustic echo control. The loudspeaker signal x is fed back to the microphone signal y. An echo control process removes this echo while ideally letting through the local near-end signal w.

Alternatively, acoustic echoes can be prevented by using an echo suppressor (AES) [4], enabling echo free half-duplex communication. If echo suppression is carried out independently at each frequency of a short-time spectral domain, a good degree of duplexity can be achieved. Recently, such an AES was introduced without the need for estimating the echo path with an FIR filter [5, 6]. The AES in [6] models the echo path with a delay and a single real-valued gain at each frequency of short-time spectra, resulting in very low computational complexity compared to a precise echo path estimation. The suppression of echoes is then implemented as a parametric Wiener filter based on a short-time power spectrum estimate of the echo.

In this paper, it is shown that the real-valued gains modeling the echo path (in the following denoted as *echo estimation filter* (EEF)) are the more biased the more noise and near-end signal is present in the microphone signal. It is also shown how the EEF can be computed without a bias, using temporal fluctuations of short-time spectra as opposed to short-time spectra.

Section 2 briefly reviews an AES using an EEF which is estimated using short-time power spectra. Then, Section 3 presents the EEF estimation based on temporal fluctuations of short-time spectra, which leads to an unbiased estimation. Remaining sections give simulations results of the proposed method and conclusions.

2. ACOUSTIC ECHO SUPPRESSOR

The room impulse response h defined in (1) can be decomposed into a direct sound, early reflections and late reflections, as illustrated in Figure 2. In the approach proposed in [6], only a global delay parameter d and a echo estimation filter (EEF) g are used to modeled the echo path in order to catch direct sound and early reflections. Late reflections are not modeled, but considered by time-smoothing of the echo suppression filter (6). The microphone signal y can thus be expressed by:

$$y[n] = g * x[n-d] + w[n]\,. \tag{2}$$

As illustrated in Figure 3, short-time discrete Fourier transform (STFT) spectra are computed from the loudspeaker and microphone signals. The delay d between the STFT windows applied to the loudspeaker signal is chosen such that most of the energy of the echo path's impulse response is captured. The STFT-domain representation of (2) is given by

$$Y[k,m] = G[k,m]X_d[k,m] + W[k,m], \tag{3}$$

where k is the block time index and m denotes the frequency index. $X_d[k,m]$ is the STFT-domain correspondence of the delayed loudspeaker signal $x[n-d]$.

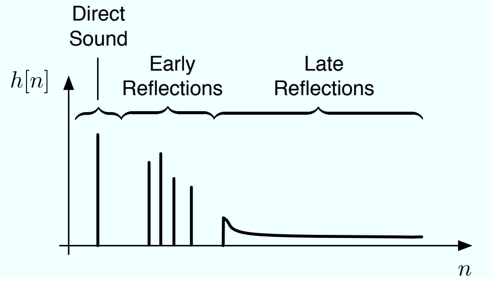

Fig. 2. The structure of a typical room impulse response.

A straightforward solution for computing the EEF $|G[k,m]|^2$ results from the signal model (3). Assuming that the near-end is silent, (3) implies that the EEF can be estimated by

$$|\hat{G}_{\text{biased}}[k,m]|^2 = \frac{\mathrm{E}\{|X_d[k,m]|^2|Y[k,m]|^2\}}{\mathrm{E}\{|X_d[k,m]|^2|X_d[k,m]|^2\}}. \tag{4}$$

Note the addition of the term *biased* in the subscript, since it will be shown later that (4) represents a biased estimator for $|G[k,m]|^2$ in case of active near-end signals.

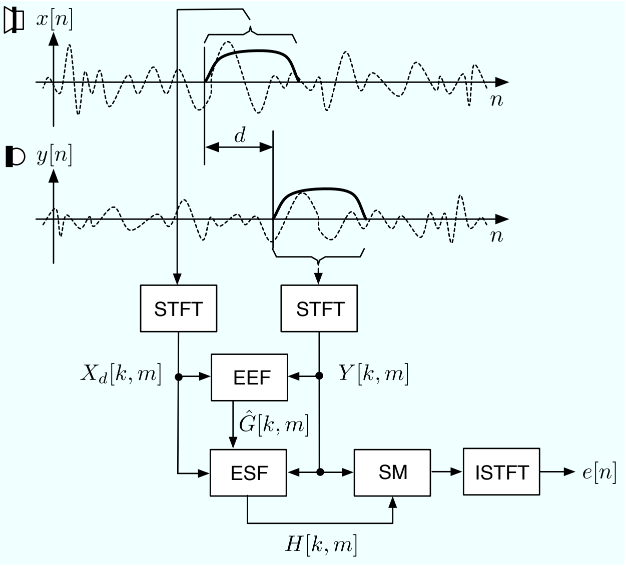

Fig. 3. Basic block diagram of a short-time spectral domain acoustic echo suppressor. STFT, ISTFT, EEF, ESF and SM stand for short-time Fourier transform, its inverse, echo estimation filter, echo suppression filter and spectral modification respectively.

For obtaining an approximate echo power spectrum, the estimated delay and EEF are applied to the loudspeaker signal power spectrum.

$$|\hat{Y}[k,m]|^2 = |\hat{G}_{\text{biased}}[k,m]|^2|X_d[k,m]|^2. \tag{5}$$

Using the echo power spectrum estimate, an echo suppression filter (ESF) is computed,

$$H[k,m] = \left(\frac{|Y[k,m]|^\alpha - \beta|\hat{Y}[k,m]|^\alpha}{|Y[k,m]|^\alpha}\right)^{\frac{1}{\gamma}}, \tag{6}$$

where α, β, and γ are filter design parameters. $H[k,m]$ is finally applied to the microphone signal spectrum to remove the echo,

$$E[k,m] = H[k,m]Y[k,m], \tag{7}$$

where $E[k,m]$ is the echo free output signal.

3. ECHO ESTIMATION FILTER

The computation of the echo estimation filter (EEF) is a crucial part of acoustic echo suppression since a suitable echo power spectrum estimation (5) depends on it. However, in the previous section, the computation of the EEF was done under the assumption that the near-end was completely silent. This assumption is not feasible since in reality background stationary noise is always present.

Assuming that the near-end signal $W[k,m]$ and the loudspeaker signal $X_d[k,m]$ are uncorrelated, it follows from (3) that $|\hat{G}_{\text{biased}}[k,m]|^2$ according to (4) gives

$$|\hat{G}_{\text{biased}}[k,m]|^2 = |G[k,m]|^2 + \frac{\mathrm{E}\{|W[k,m]|^2\}}{\mathrm{E}\{|X_d[k,m]|^2\}}. \qquad (8)$$

Obviously, any non-negligible near-end signal included in the microphone signal $Y[k,m]$ leads to a bias in the estimate of $|G[k,m]|^2$. The biased EEF leads to too large estimates of the echo power in the spectrum. From (8) it follows that this effect is especially prominent in case of high levels of the near-end signal $W[k,m]$, which makes the use of a double talk control necessary.

In the following, we propose a method to compute the EEF without a bias. This is achieved by estimating $|G[k,m]|^2$ based on temporal fluctuations of the power spectra of the loudspeaker and microphone signals. The temporal fluctuations of the power spectra are computed according to

$$\tilde{Y}[k,m] = |Y[k,m]|^2 - \mathrm{E}\big\{|Y[k,m]|^2\big\} \qquad (9)$$

$$\tilde{X}_d[k,m] = |X_d[k,m]|^2 - \mathrm{E}\big\{|X_d[k,m]|^2\big\}\,, \qquad (10)$$

where $\mathrm{E}\big\{|Y[k,m]|^2\big\}$ and $\mathrm{E}\big\{|X_d[k,m]|^2\big\}$ are estimated using a single-pole averaging scheme with a time constant τ, which determines the exponential decay of the estimation window. The estimation of the EEF is then performed analogously to (4), but based on the fluctuating spectra of the loudspeaker and the microphone:

$$|\hat{G}[k,m]|^2 = \frac{\mathrm{E}\Big\{\tilde{X}_d[k,m]\tilde{Y}[k,m]\Big\}}{\mathrm{E}\Big\{\tilde{X}_d[k,m]\tilde{X}_d[k,m]\Big\}}. \qquad (11)$$

For simplicity of notation, the time and frequency indices k and m are omitted in the following. The numerator of (11) is equal to

$$\mathrm{E}\Big\{\tilde{X}_d\tilde{Y}\Big\} = \\ \mathrm{E}\big\{\big(|Y|^2 - \mathrm{E}\{|Y|^2\}\big)\big(|X_d|^2 - \mathrm{E}\{|X_d|^2\}\big)\big\}\,. \qquad (12)$$

Assuming that the near-end, $|W|^2$, and the loudspeaker, $|X_d|^2$, contributions are orthogonal, the microphone power spectrum can be obtained as

$$|Y|^2 = |G|^2|X_d|^2 + |W|^2, \qquad (13)$$

and, thus, (12) can be expressed as

$$\mathrm{E}\Big\{\tilde{X}_d\tilde{Y}\Big\} = |G|^2\Big(\mathrm{E}\{|X_d|^4\} - \big(\mathrm{E}\{|X_d|^2\}\big)^2\Big)\,. \qquad (14)$$

In the same way the denominator can be derived to be

$$\mathrm{E}\Big\{\tilde{X}_d\tilde{X}_d\Big\} = \mathrm{E}\{|X_d|^4\} - \big(\mathrm{E}\{|X_d|^2\}\big)^2\,.$$

Thus, (11) yields an un-biased echo estimation filter, i.e. $|\hat{G}|^2 = |G|^2$.

4. SIMULATIONS AND EVALUATIONS

The proposed AES is implemented using a discrete short-time Fourier transform (STFT). The presented simulations use a sampling rate of 16 kHz and an STFT window length of 512 (FFT size). Successive frames are defined with an overlap of 50%. In order to reduce computational complexity, the processing is not carried out on each STFT frequency bin separately. The uniformly spaced spectral coefficients are grouped into a number of non-overlapping partitions, similar as described in [7]. The number of partitions is 16.

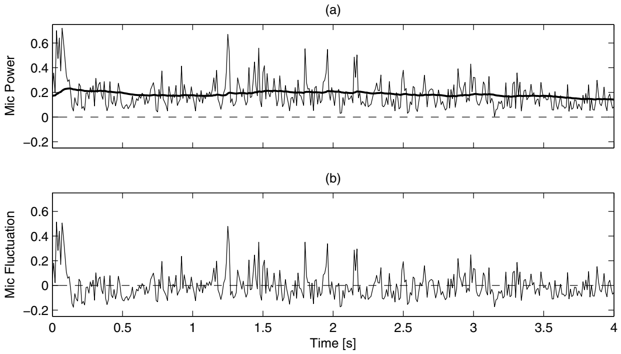

Fig. 4. Part (a) shows the short-time power spectral values $|Y[k,m]|^2$ and the corresponding mean $\mathrm{E}\big\{|Y[k,m]|^2\big\}$ (bold) of a noisy speech signal at a frequency of 1000 Hz. Part (b) shows the corresponding temporal fluctuations $\tilde{Y}[k,m]$.

In order to cancel out the contribution of the near-end signal (undesired bias) in the echo estimation filter (EEF), temporal fluctuations of power spectra, as defined in (9) and (10), are used for computing the EEF. Panel (a) of Figure 4 shows an example of microphone short-time power spectral values, $|Y[k,m]|^2$, at 1000 Hz and the corresponding temporal mean, $\mathrm{E}\big\{|Y[k,m]|^2\big\}$, estimated for a female speech signal with Gaussian noise with an SNR of 12 dB. The choice of the time constant τ for the mean computation is an important parameter and has been chosen equal to 300 ms. Panel (b) of Figure 4 shows the corresponding temporal fluctuations $\tilde{Y}[k,m]$.

The next simulations aim at showing that the EEF, based on temporal fluctuations, results in less bias than the same filter computed on the power directly. The simulations consider a far-end female speech signal with additive Gaussian noise with an SNR of 24 dB. The microphone signal contains the echo and near-end Gaussian noise with three different SNRs: 6 dB, 12 dB, and 24 dB. A measured room impulse response h of length 64 ms is used to produce realistic echo.

For a specific partition at 1000 Hz, the dotted line in the three panels of Figure 5 shows the EEF $G[k,m]$, estimated without noise in the far-end speech and microphone signal as a reference. The biased estimate $\hat{G}_{\text{biased}}[k,m]$ and the unbiased estimate $\hat{G}[k,m]$ are both shown with dashed and solid lines, respectively. Panel (a) represents the corresponding estimates with a microphone SNR of 24 dB, whereas Panels (b) and (c) with SNR 12 dB and 6 dB, respectively. The data indicate that the microphone SNR has an impact on the biased estimate, as expected from (8). The worse the microphone SNR is, the more biased is the EEF $\hat{G}_{\text{biased}}[k,m]$. As expected, the unbiased EEF estimate $\hat{G}[k,m]$ depends less on the microphone SNR. After convergence, its value is very similar to the reference value (dotted line).

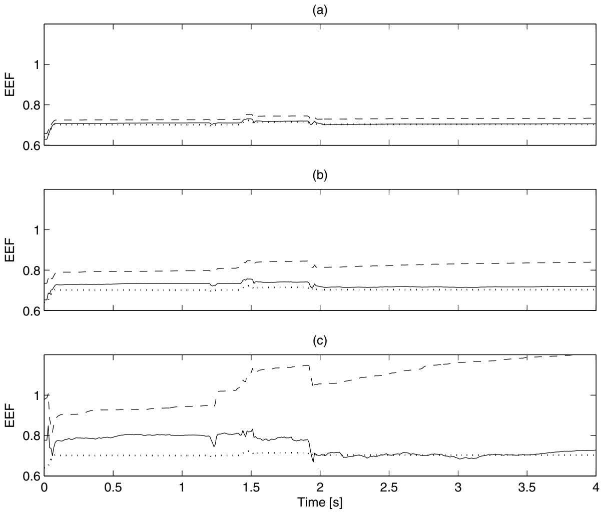

Fig. 5. The reference EEF $G[k,m]$ (dotted), its biased $\hat{G}_{\text{biased}}[k,m]$ (dashed) and unbiased $\hat{G}[k,m]$ (solid) estimates are shown for a frequency of 1000 Hz for several microphone SNRs: 24 dB in Panel (a), 12 dB in Panel (b), and 6 dB in Panel (c).

Figure 6 shows the same data, not only for the partition at 1000 Hz, but for the partitions at all frequencies. The left three panels show the biased EEF estimates, while the right three panels show the unbiased EEF estimates for the same different microphone SNRs: 24 dB, 12 dB, and 6 dB, respectively. While the unbiased EEF estimates $\hat{G}[k,m]$ are similar for all SNR conditions, the biased EEF estimates are the more impaired the lower the SNR is.

The simulation results show that stationary noise in the microphone signal leads to a biased EEF $\hat{G}_{\text{biased}}[k,m]$. The proposed method eliminates this bias, even in case of low microphone SNR, and may also allow estimating suitable EEF in double talk situations, which is part of further investigations.

5. CONCLUSIONS

An improvement for an acoustic echo suppressor, based on a short-time spectral echo estimation filter (EEF), is proposed in this paper. This type of echo control estimates the echo power spectrum without the need for an accurate echo path model, which enables low computational complexity and high robustness. It is shown that estimation of EEF using short-time power spectra results in a bias related to microphone signal SNR. The proposed EEF estimation based on temporal flucations of short-time power spectra is not biased even when the microphone signal has low SNR. The presented simulation results indicate that the proposed EEF estimation avoids bias also in realistic communication scenarios.

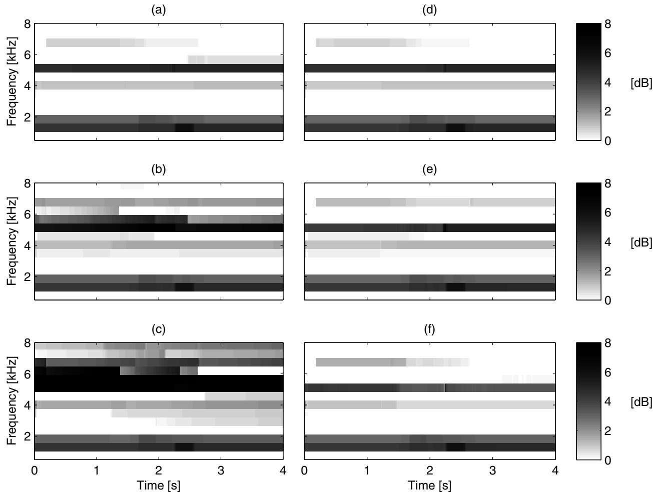

Fig. 6. Panels (a), (b), and (c) show the biased EEF estimates for the different conditions, 24 dB, 12 dB, and 6 dB, respectively. Panels (d), (e), (f) show the corresponding unbiased EEF estimates also for SNRs of 24 dB, 12 dB, and 6 dB.

6. REFERENCES

[1] G. Schmidt and E. Hänsler, *Acoustic echo and noise control: a practical approach*, Hoboken: Wiley, 2004.

[2] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. 46, pp. 497–510, Mar. 1967.

[3] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer, 2001.

[4] M. M. Sondhi and D. A. Berkeley, "Silencing echoes on the telephony network," *Proc. IEEE*, vol. 68, pp. 948–963, Aug. 1980.

[5] C. Faller and J. Chen, "Suppressing acoustic echo in a sampled auditory envelope space," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 5, pp. 1048–1062, Sept. 2005.

[6] C. Faller and C. Tournery, "Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression," in *Proc. Intl. Works. on Acoust. Echo and Noise Control (IWAENC)*, Sept. 2005.

[7] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.