ESTIMATING THE DELAY AND COLORATION EFFECT OF THE ACOUSTIC ECHO PATH FOR LOW COMPLEXITY ECHO SUPPRESSION

[1]*Christof Faller and* [1]*Christophe Tournery*

[1]{christof.faller, christophe.tournery}@epfl.ch
[1] Audiovisual Communications Laboratory, EPFL Lausanne, Switzerland

ABSTRACT

Acoustic echoes arise whenever there is acoustic coupling between a loudspeaker and a microphone. A traditional solution for eliminating the undesired echo signal is an acoustic echo canceler (AEC), which identifies the echo path between a loudspeaker and a microphone by means of an adaptive filter. The echo signal can be canceled successfully when the modeling filter approaches the true echo path. Another way to mitigate the echo effect is through echo suppression. Unlike an AEC, an acoustic echo suppressor (AES) achieves echo attenuation by means of spectral modification. In this paper, we propose an AES without a need for the computationally complex task of acoustic echo path estimation. Instead of identifying the echo path impulse response, the proposed method estimates only the properties of the echo path which are needed for effective echo suppression. These are a delay parameter and a filter mimicking the coloration effect of the echo path on the loudspeaker signal. The gain filter for the AES is computed using the estimated delay and coloration effect filter. Simulations and experiments using a real-time telecommunication system indicate that the proposed scheme is effective for suppressing echo, supports duplex communication, and is less sensitive to echo path changes than a conventional AEC.

1. INTRODUCTION

Acoustic echo control is a necessary component for a full-duplex hands-free telecommunication system to eliminate undesired echo signals that result from acoustic coupling between a loudspeaker and a microphone. Usually, an *acoustic echo canceler* (AEC) [1] is used to remove the undesired echo signal component from the microphone signal. An AEC achieves the echo removal by modeling the echo path impulse response with an adaptive finite impulse response (FIR) filter and subtracting an echo estimate from the microphone signal. It is not uncommon that an adaptive filter with a length of 50 – 300 milliseconds needs to be considered, which makes an AEC highly computationally expensive. If an AEC is used, usually also an *acoustic echo suppressor* (AES) is used (in series after the AEC) to remove residual echoes which occur due to the constantly changing echo paths or when sudden echo path changes occur. Also, often a *noise suppressor* (NS) is applied for removing stationary noise from the microphone signal.

Recently, systems have been proposed which do not employ an AEC, but do all echo removal using an AES [2–4]. However, these systems have still high complexity [2] or do not estimate a measure which is signal independent [4], such as the echo path. We are proposing a scheme for AES which has as low complexity as [3, 4] but better performance and higher robustness because it estimates signal independent physical properties of the echo path.

2. ACOUSTIC ECHO SUPPRESSOR (AES)

Unlike AEC, an AES achieves echo attenuation through manipulating the magnitude spectrum of the microphone signal in the frequency domain, while leaving the phase spectrum untouched. For noise suppression, a widely adopted spectral manipulation algorithm is the parametric Wiener filter (or sometimes called spectral subtraction [5]). If $|\hat{Y}(i,k)|$ denotes an estimate of the magnitude spectrum of the echo signal with frequency index i and time index k, a parametric Wiener filter based echo suppression algorithm can be expressed as

$$e(n) = \mathrm{F}^{-1}[G(i,k)Y(i,k)]\,, \tag{1}$$

where e(n) is the echo-suppressed outgoing signal, $Y(i,k)$ is the short time spectrum of the microphone signal, $\mathrm{F}^{-1}[\cdot]$ denotes the inverse Fourier transform, and

$$G(i,k) = \left[\frac{\max(|Y(i,k)|^\alpha - \beta|\hat{Y}(i,k)|^\alpha, 0)}{|Y(i,k)|^\alpha}\right]^{\frac{1}{\alpha}} \tag{2}$$

is a Wiener gain filter, where α and β are design parameters to control the echo suppression performance [6]. If the echo is under-estimated, $\beta > 1$ is used, and $\beta < 1$ if it is over-estimated. In this paper, we are proposing a highly computationally efficient method for obtaining $|\hat{Y}(i,k)|$.

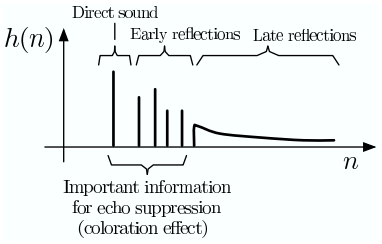

Figure 1: *The structure of a typical room impulse response (acoustic echo path).*

3. THE PROPOSED AES

The difference between the proposed scheme and other approaches is that not the acoustic echo path is estimated, but merely a global delay parameter and a filter characterizing the coloration effect of (the early part of) the acoustic echo path. This representation (delay and coloration effect filter) is largely insensitive to acoustic echo path changes and is thus more robust than conventional methods which estimate the acoustic echo path. Additionally, the computational complexity is much lower as will be explained.

In audio processing, with *coloration* it is usually meant that some frequency ranges are attenuated while other frequencies are not attenuated or amplified. This is called coloration because such audio signals are perceived as being "colored". For echo suppression, it is important to know which frequencies are attenuated, not modified, or amplified by the echo path. Given this information and the estimated delay, the echo signal can be suppressed. A room impulse response (the acoustic echo path) usually features the direct sound (sound that travels directly from the loudspeaker to the microphone), followed by a few early reflections and a tail consisting of late reflections with high density. Figure 1 illustrates the structure of a typical room impulse response. The direct sound and the early reflections have a coloration effect on the audio signal. The densely spaced late reflections do not or hardly color the signal. Thus, for obtaining the information necessary for an effective echo suppression gain filter it is enough to only consider the direct sound and early reflections.

The proposed scheme is illustrated in Figure 2. Short time Fourier transform (STFT) spectra are computed from the loudspeaker and microphone signals. A delay d between the STFTs applied to microphone and loudspeaker signal is chosen such that most of the effect of the echo path impulse response is captured. Then, the real-valued coloration effect filter, $G_V(i,k)$, mimicking the effect of the early echo path, is estimated. For obtaining an approximate echo magnitude spectrum, the estimated delay and coloration effect filter are applied to the loudspeaker signal spectra,

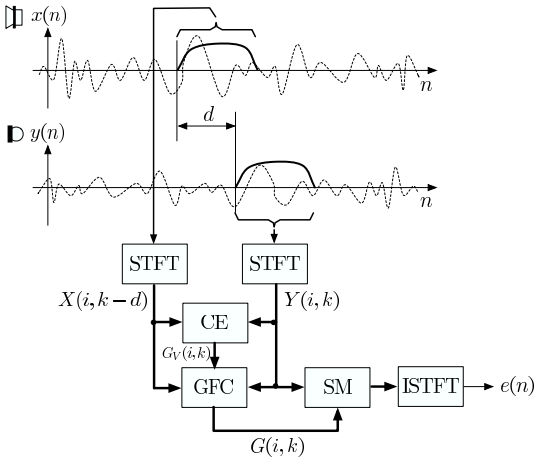

Figure 2: *Block diagram of the proposed acoustic echo suppression algorithm. STFT, ISTFT, CE, GFC, and SM stand for short time Fourier transform, its inverse, coloration effect estimation, gain filter computation, and spectral modification, respectively.*

$$|\hat{Y}(i,k)| = G_V(i,k)|X_d(i,k)|\,, \tag{3}$$

where d indicates that the spectrum is computed with a waveform that is delayed by d samples. Underestimation of the echo signal magnitude spectrum due to ignoring the late reflections can be compensated for by using a $\beta > 1$ (2). Note that this is not a precise echo spectrum or magnitude spectrum estimate. But it contains the information necessary for applying echo suppression, i.e. (1) with (2).

4. ADAPTIVE ESTIMATION OF THE DELAY AND COLORATION EFFECT FILTER

Delay estimation is often also an issue with conventional AECs since the delay of the audio input and output in personal computers is often not known and depends on the specific audio devices which are used. In order that only as many filter taps as necessary can be used with an AEC, the delay of the audio input and output has to be known. Delay estimation algorithms, similar as are used for a conventional AEC can be applied for computing d for the proposed algorithm. We are estimating the delay by computing the temporal envelopes of the microphone and loudspeaker signals. Then, the cross-correlation between these temporal envelopes is computed and the delay estimate is chosen to be the lag value at which the cross-correlation function is maximized.

The gain filter is computed as the magnitude of the least squares estimator

$$G_V(i,k) = \left| \frac{\mathrm{E}\{X_d^*(i,k)Y(i,k)\}}{\mathrm{E}\{X_d^*(i,k)X_d(i,k)\}} \right|\,, \tag{4}$$

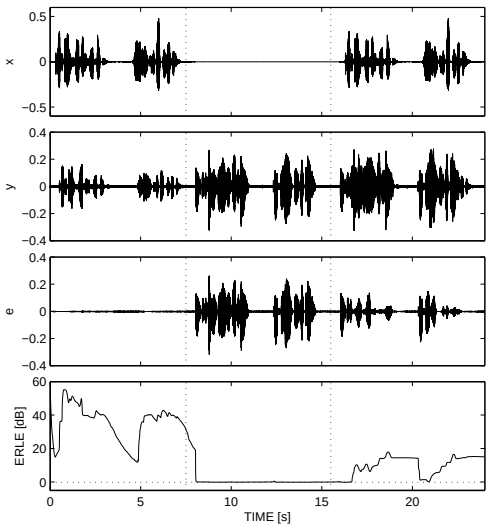

Figure 3: *The loudspeaker signal x, microphone signal y, AES output signal e, and ERLE are shown.*

where * denotes complex conjugate. Since the acoustic echo path is likely to vary in time, $G_V(i,k)$ is estimated iteratively by

$$G_V(i,k) = \frac{a_{12}(i,k)}{a_{22}(i,k)}\,, \tag{5}$$

where

$$\begin{aligned} a_{12}(i,k) &= \epsilon|X_d^*(i,k)Y(i,k)| + (1-\epsilon)a_{12}(i,k-1) \\ a_{22}(i,k) &= \epsilon X_d^*(i,k)X_d(i,k) + (1-\epsilon)a_{22}(i,k-1)\,, \end{aligned}$$

and $\epsilon \in [0,1]$ determines the time-constant of the exponentially decaying estimation window

$$T = \frac{1}{\epsilon f_s}\,, \tag{6}$$

where f_s denotes the STFT spectrum sampling frequency. We use $T = 1.5$ s.
To prevent that during periods of doubletalk the coloration effect filter $G_V(i,k)$ diverges, we use two coloration effect filters, similarly as two echo path models have been used for conventional AEC [7].

5. SIMULATIONS AND EVALUATIONS

The audio signals are processed in blocks of length 10 ms. The simulations are carried out with 16 kHz sampling frequency, for which blocks of 160 samples are processed at a time. A FFT of size 512 is used with a sine window (analysis and synthesis) of length 320 with 50% window hop size. The computational complexity of the proposed scheme is much lower than a conventional AEC, since only the real-valued coloration effect filter values $G_V(i,k)$ need to be estimated as opposed to the echo path with many more parameters (filter taps).
A dialogue sequence is used, starting with far-end only speech (loudspeaker signal), followed by near-end only speech, and concluding with far-end and near-end speech simultaneously (doubletalk). The signal-to-noise ratio of the microphone signal is 20 dB. Measured impulse responses with 4096 taps are used for generating the loudspeaker and microphone signals. The impulse responses were measured using a computer-based desktop audio system.

Echo suppression and doubletalk performance: The top two panels of Figure 3 show the loudspeaker and microphone signals, $x(n)$ and $y(n)$, resulting from the described dialogue sequence. The vertical dotted lines separate the three parts of the simulation: Far-end only speech, near-end only speech, and doubletalk. The bottom two panels show the echo suppressed AES output signal $e(n)$ and the echo return loss enhancement (ERLE) in dB, respectively. The ERLE is computed as a short-time estimate of

$$\text{ERLE} = 10\log_{10}\frac{\mathrm{E}\{y^2(n)\}}{\mathrm{E}\{e^2(n)\}}\,. \tag{7}$$

The ERLE and $e(n)$ imply that during far-end only speech the echo is instantly suppressed. The instant suppression is due to the initial values for $G_V(i,k)$ which we are setting such that the echo is overestimated initially and thus suppressed. The near-end only speech gets through unimpaired. The doubletalk is partially suppressed as indicated by $e(n)$ and the ERLE in the figure.

Echo magnitude spectrum estimation: The top two panels of Figure 4 show the magnitude spectrum of the loudspeaker and microphone signals, $|X_d(i,k)|$ and $|Y(i,k)|$, respectively. The third panel from the top shows the estimated echo signal magnitude spectrum $|\hat{Y}(i,k)|$ and the bottom panel shows the estimated coloration effect filter $G_V(i,k)$. Visual inspection of $|\hat{Y}(i,k)|$ during far-end only speech indicates that it is similar to $|Y(i,k)|$, as desired. $G_V(i,k)$ quickly converges during the far-end only speech and is only little affected during near-end only speech and doubletalk time periods.

Robustness: We repeated simulations using the same dialogue sequence as before. To test the robustness of the proposed algorithm we modified the loudspeaker and microphone signals.
In a first experiment, we toggled between two echo path impulse responses every second. The two echo path responses used were measured for the two loudspeakers of a desktop stereo system. The resulting ERLE is shown in Figure 5. Comparison of this result with the bottom panel of Figure 3 indicates that the performance of the

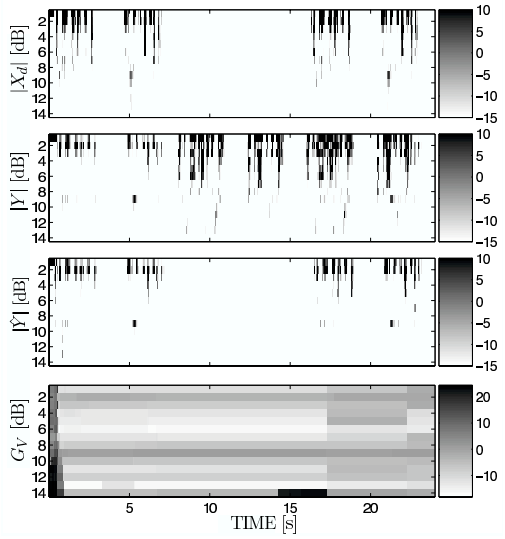

Figure 4: *The magnitude spectra of the loudspeaker signal $|X_d|$, microphone signal $|Y|$, and the estimated echo signal $|\hat{Y}|$ are shown. The estimated gain filter G_V is also shown. The number of frequency bands has been reduced by averaging.*

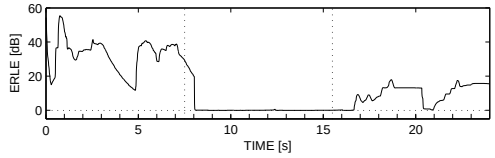

Figure 5: *The ERLE for the same simulation as shown in Figure 3, but simulating echo path changes every second. The ticks on the x-axis indicate the time instants when the echo path is changed.*

proposed algorithm is very similar for the case when echo path changes occur, indicating high robustness.

In a second experiment, we simulated non-linear distortion of the loudspeaker. The loudspeaker signal was processed as $\tilde{x}(n) = \text{sign}(x(n))\text{abs}(x(n))^{1.4}$, where $\text{sign}(a) = 1$ for $a \geq 0$ and $\text{sign}(a) = -1$ for $a < 0$. The range of the loudspeaker signal was $[-1,\ 1]$. The non-distorted loudspeaker signal is given to the AES algorithm, while the echo signal is computed using the distorted loudspeaker signal. Figure 6 shows the ERLE for this simulation. Again, comparing this result with the bottom panel of Figure 3 implies that the proposed algorithm is largely insensitive to this type of non-linear signal distortion.

Real-time implementation: We implemented a library of the proposed algorithm compatible with all major operating systems (Linux, MacOS X, Windows). Sampling rates up to 48 kHz are supported. For 16 kHz sampling rate the proposed AES consumes only about one percent

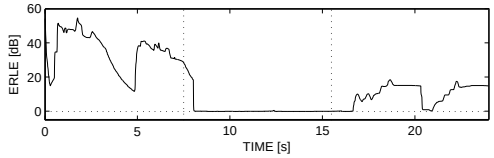

Figure 6: *The ERLE for the same simulation as shown in Figure 3, but simulating non-linear loudspeaker distortions.*

of the processing power on a 1.6 GHz Pentium M processor. Informal testing with the real-time conferencing system indicates effectivity of the proposed algorithm in terms of echo suppression, doubletalk performance, and robustness.

6. CONCLUSIONS

We proposed a low complexity acoustic echo control algorithm. As opposed to identifying the acoustic echo path, only an overall delay parameter and a gain filter mimicking the coloration effect of the echo path are estimated. Given this, a spectral modification algorithm is applied for removing the acoustic echo signal from the microphone signal.

Numerical simulations and testing with a real-time implementation indicate that the proposed algorithm effectively suppresses echo and that it is robust against non-linearities and minor echo path changes. The computational complexity of the algorithm is very low.

7. REFERENCES

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

[2] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2001.

[3] C. Faller, "Perceptually motivated low complexity acoustic echo control," in *Preprint 114th Conv. Aud. Eng. Soc.*, Mar. 2003.

[4] C. Faller and J. Chen, "Suppressing acoustic echo in a sampled auditory envelope space," *IEEE Trans. on Speech and Audio Proc.*, Aug. 2003, (submitted Aug. 2003, accepted).

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE trans. Acoust. Speech Sig. Processing*, vol. 27, no. 2, pp. 113–120, Nov. 1979.

[6] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[7] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceler with two echo path models," *IEEE trans. on Communications*, vol. 25, no. 6, pp. 589–595, June 1977.