STEREO ACOUSTIC ECHO CONTROL USING A SIMPLIFIED ECHO PATH MODEL

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

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

ABSTRACT

In handsfree tele- or video-communication, acoustic echoes arise due to the coupling between the loudspeakers and microphones. It is much more challenging to remove the undesired acoustic echoes for stereo or multi-channel tele-communication systems than for mono systems due to the non-uniqueness problem. While non-uniqueness can be prevented by introducing independent distortions into the left and right loudspeaker signals, stereo echo cancellation is more challenging in terms of convergence speed and computational complexity than mono echo cancellation. The proposed stereo echo control algorithm circumvents the non-uniqueness problem by using simplified echo path models consisting of delays and short-time spectral modification. It is shown that for reasonably symmetric systems the left and right echo path models are similar enough that a single echo path model can be used for estimating the total echo power spectrum and a gain filter for removing the echo from the microphone channels. The proposed algorithm is also applicable to multi-channel systems and the computational complexity is very low.

1. INTRODUCTION

While from a theoretical point of view an acoustic echo canceler (AEC) can perfectly remove the echo without introducing distortions, in practice, often problems arise due to non-linearities in the effective transfer function from the loudspeaker signals to the microphone signals. Such non-linearities are due to time drift when different clocks are used for audio input and output (e.g. VoIP using a web-cam), overdriving of small loudspeakers (handsfree portable devices), auto gain control, flaws in the analog circuitry, etc.

It has recently been shown that not only an acoustic echo canceler (AEC) can provide echo control for duplex communication, but also that an acoustic echo suppressor (AES), operating in a short-time spectral domain, can provide a high degree of duplex capability. The above mentioned problems, indicating the weaknesses of AEC in practical situations, may suggest that an AES may be a viable alternative to an AEC despite of its potential for artifacts during doubletalk. In [1] AEC and AES were compared in terms of speech quality during doubletalk, indicating a relatively small quality difference under non-ideal conditions.

In [2] echo removal by means of spectral modification was proposed for improving the robustness of a frequency domain acoustic echo canceler. An AES without a need for estimating the acoustic echo path impulse response was proposed in [3]. Recently, we proposed an improved AES using a simplified echo path model [4], aiming at low computational complexity and high robustness. The echo path is modeled by an overall delay parameter and a coloration effect filter which captures the effect of the echo path in terms of short-time spectral modification.

Since acoustic echo cancellation is more difficult for stereo or multi-channel tele-communication [5], AES may also be a viable alternative to AEC for this scenario. In this paper, we are proposing an AES similar to [4], but which is applicable for removing acoustic echoes for stereo and multi-channel tele-communication. To avoid the non-uniqueness problem, we are assuming that the tele-communication clients use a reasonably symmetric loudspeaker and microphone setup. As will be shown, the symmetry does not have to be very strict.

The paper is organized as follows. Section 2 reviews the previously proposed AES algorithm which also forms the basis of the proposed multi-channel AES algorithm. The proposed AES algorithm, applicable for multi-channel audio signals, is described in Section 3. The results of a number of simulations are presented in Section 4. The conclusions are presented in Section 5.

2. ACOUSTIC ECHO CONTROL USING A SIMPLIFIED ECHO PATH MODEL

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 [6]). 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 sup-

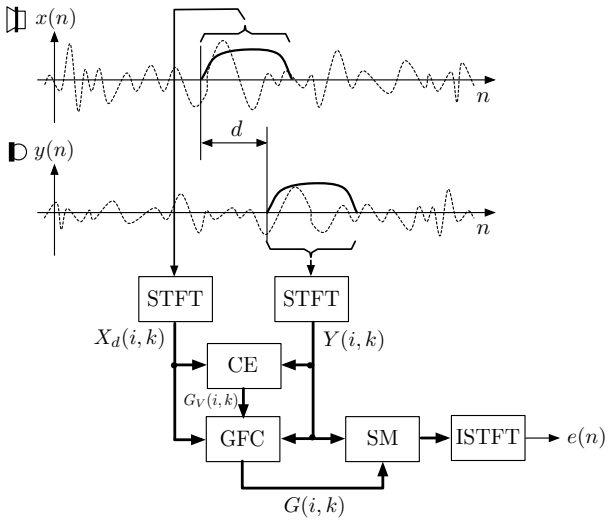

Figure 1: *Acoustic echo suppressor (AES) using a simplified echo path model.*

pression 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 [7]. If the echo is under-estimated, $\beta > 1$ is used, and $\beta < 1$ if it is over-estimated.

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 [4]. 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 significantly lower since much less parameters need to be estimated.

The described AES is illustrated in Figure 1. 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. The CE block in the figure estimates a real-valued coloration effect filter, $G_V(i,k)$, mimicking the effect of the early echo path. For obtaining an approximate echo magnitude spectrum, the estimated delay and coloration effect filter are applied to the loudspeaker signal spectra,

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

where d indicates that the spectrum is computed with a waveform that is delayed by d samples. Time smoothing of the gain filter can compensate for the fact that the late part of the echo path is ignored. Note that (3) is not a precise echo spectrum or magnitude spectrum estimate, but it contains the information necessary for gain filter computation (2) (GFC block).

The coloration effect filter is computed as the magnitude of the least squares estimator

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

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 [8]. The background filter, $G_V(i,k)$, is always adapted and a static foreground filter, $H_V(i,k)$, is used. Whenever there is confidence that $G_V(i,k)$ is a good estimate it is copied to $H_V(i,k)$, which is used for echo estimation.

3. MULTI-CHANNEL ACOUSTIC ECHO CONTROL

We observed that the coloration correction filters corresponding to the echo paths of two loudspeakers in a desktop stereo system are quite similar for the left and right loudspeakers (if there is reasonable symmetry, i.e. if the microphone is not placed much closer to one loudspeaker than the other). We did this experiment with an omnidirectional microphone. If directional microphones are

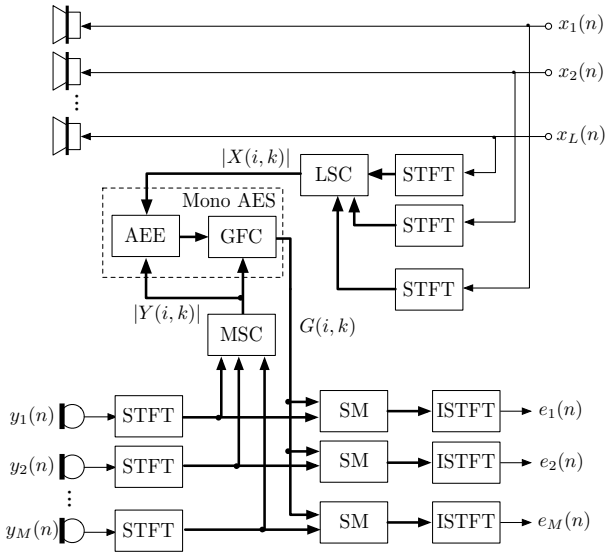

Figure 2: *Multi-channel acoustic echo suppressor (AES) using a single simplified echo path model.*

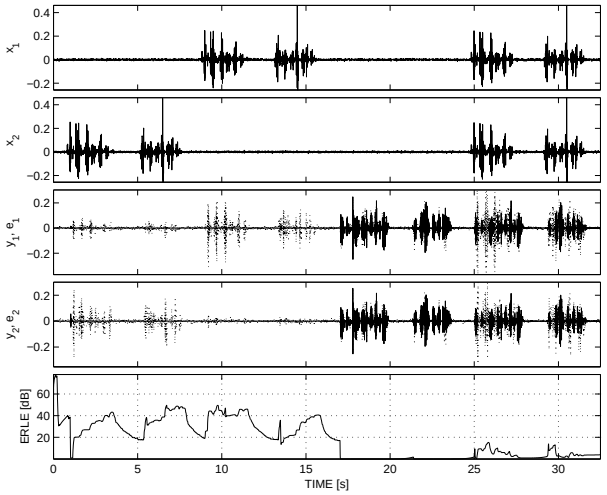

Figure 3: *Shown are: Left and right loudspeaker signals (x_1 and x_2), left and right microphone signals (y_1, y_2, dotted), left and right echo suppressed signals (e_1, e_2, solid), and the total ERLE in dB.*

used, the coloration effect is still similar, but the overall gain depends on the directions from which the direct sound and strongest reflections arrive at the microphone. Often stereo microphones are designed such that the sum of left and right has an omnidirectional property (i.e. the gain of the left and right sum does not depend on direction). This omnidirectional property of the sum signal motivated us to combine the left and right microphone signal spectra and treat the combined spectrum the same as a single microphone spectra for gain filter $G(i,k)$ computation. Spinning this thought further, we also tried to combine the loudspeaker signals to a single signal for gain filter computation. In the end, we had a system effectively using mono gain filter computation applicable for stereo and multi-channel AES and thus avoiding the non-uniqueness problem. In the following, we are describing this process in detail.

Figure 2 shows how a scheme for mono AES is extended for multi-channel acoustic echo suppression. Note that the AEE block in the figure corresponds to a method for estimating an echo signal spectrum, e.g. by applying a delay and coloration correction to the loudspeaker signal, i.e. (3). A loudspeaker signal combiner (LSC block in the figure) combines the loudspeaker signal spectra and generates a "combined" magnitude spectrum $|X(i,k)|$. The loudspeaker signals are combined as

$$|X(i,k)| = \left(\sum_{l=1}^{L} g_{x_l} |X_l(i,k)|^{\theta} \right)^{\frac{1}{\theta}} , \qquad (7)$$

where θ controls the combination process and g_{x_l} are weighting factors for each signal. We use $\theta = 2$ and $g_{x_l} = 1$.

Similarly, a microphone signal combiner (MSC) combines the microphone signal spectra,

$$|Y(i,k)| = \left(\sum_{m=1}^{M} g_{y_m} |Y_m(i,k)|^{\lambda} \right)^{\frac{1}{\lambda}} , \qquad (8)$$

where λ controls the combination process and g_{y_m} are weighting factors. We use $\lambda = 2$ and $g_{y_m} = 1$.

Given the combined magnitude spectra, $|X(i,k)|$ and $|Y(i,k)|$, the gain filter, $G(i,k)$, is computed similarly as in the mono AES case, as illustrated in Figure 2. That is, the echo magnitude spectrum $|\hat{Y}(i,k)|$ is estimated and the gain filter $G(i,k)$ (2) is computed. Spectral modification is then applied to each of the microphone signals $1 \leq m \leq M$ individually, using the *same* gain filter $G(i,k)$,

$$E_m(i,k) = G(i,k) Y_m(i,k) . \qquad (9)$$

The echo suppressed output signals $e_m(n)$ are obtained by applying the inverse STFT with overlap add to $E_m(i,k)$.

4. SIMULATIONS

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

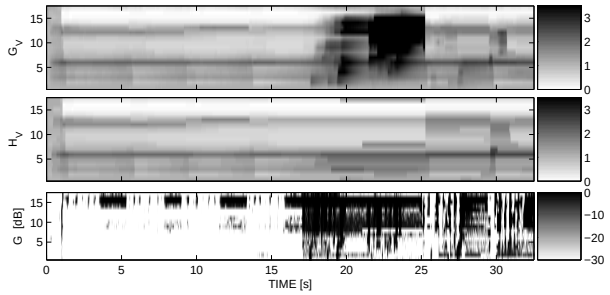

Figure 4: *Shown are the coloration correction filter, G_V, as a function of time and frequency (top panel), the foreground filter, H_V, that is actually used (middle panel), and the gain filter G used for echo removal.*

with many more parameters (filter taps). For further reducing complexity, we are combining and jointly processing STFT spectral coefficients, mimicking the frequency resolution of the auditory system. On a 1.6 GHz Pentium M based laptop computer, the proposed algorithm needs $1.2\,\%$ processing power for two-channel stereo and a sampling rate of 16 kHz.

We use $\alpha = 2$ and $\beta = 2.5$ in (2). Due to the relatively large β parameter, the system becomes less sensitive with respect to asymmetry of the loudspeaker setup. Extensive testing with our real-time stereo tele-conferencing system implies that indeed the symmetry requirement is only weak, e.g. it is not a problem if the microphone is twice as close to one loudspeaker than the other one.

A dialogue sequence is used, starting with far-end only speech on the right (right loudspeaker signal), followed by far-end only speech on the left (left 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 loudspeaker and microphone signals are 20 dB. Impulse responses measured with a stereo desktop system and coincident pair stereo microphone with 4096 taps are used for generating the loudspeaker and microphone signals.

The top four panels of Figure 3 show the loudspeaker signals, $x_1(n)$ and $x_2(n)$ and the microphone signals, $y_1(n)$ and $y_2(n)$ (indicated as dotted lines), resulting from the described dialogue sequence. The third and fourth panel also show the echo suppressed stereo AES output signals, $e_1(n)$ and $e_2(n)$ (indicated as solid lines). The bottom panel shows the total echo return loss enhancement (ERLE) in dB.

The ERLE, $e_1(n)$, and $e_2(n)$ imply that during far-end only speech the echo is instantly suppressed. The instant suppression in the beginning is due to the initial values for $G_V(i,k)$ which are such that the echo is overestimated initially and thus suppressed. The near-end only speech gets through unimpaired. When the far-end speech changes from left to right, echo stays fully suppressed as desired. The doubletalk is let through as indicated by $e_1(n)$, $e_2(n)$, and the ERLE in the figure.

Figure 4 shows the coloration effect filter $G_V(i,k)$ for the same simulation. $G_V(i,k)$ does not notably change when during the first 15 s the far-end speech moves from right to left, indicating, as assumed, that $G_V(i,k)$ does not strongly depend on whether sound is emitted from the left or right loudspeaker. The foreground filter, H_V, actually used for echo estimation stays relatively constant over the whole simulation, also during doubletalk, indicating effectivity of the doubletalk control. The gain filter, $G(i,k)$, used for echo removal is also shown in the figure.

5. CONCLUSIONS

Extension of a previously proposed algorithm for acoustic echo suppression for stereo and multi-channel teleconferencing was proposed. For computing the gain filter for echo removal, single combined loudspeaker and microphone spectra are used. The so-obtained gain filter is applied to each of the microphone signals. Simulations and real-time implementation indicate effectivity of the proposed algorithm.

6. REFERENCES

[1] F. Wallin and C. Faller, "Perceptual quality of hybrid echo canceler/suppressor," in *Proc. ICASSP*, May 2004.

[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 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.

[4] 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.

[5] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, pp. 148–151, Aug. 1995.

[6] 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.

[7] 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.

[8] 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.