ICC Systems Require Multichannel Acoustic Echo Cancellation: How to Perform Efficient Residual Echo Suppression

Jan Franzen, Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig,

Schleinitzstr. 22, 38106 Braunschweig, Germany, Email: {franzen, fingscheidt}@ifn.ing.tu-bs.de

Abstract

In-car communication (ICC) systems reproduce amplified speech from the car cabin in the car cabin to support speech communication between the passengers in noisy conditions. The main component of an ICC system is an acoustic echo (or feedback) cancellation (AEC). To address the typically remaining residual echo, a postfilter for residual echo suppression is usually applied afterwards. Facing the scenario of passengers having a conversation while stereo music is played from the audio player or FM radio, we show an efficient residual echo suppression postfilter that can be used in combination with the stereo-channel Kalman Filter AEC, thus allowing to simultaneously cancel the echoes stemming from the audio player or FM radio.

Introduction

In-car communication (ICC) systems support the speech communication between rear and front seat passengers in a car. They typically use the car's existing microphones to acquire speech and send it to the loudspeakers at the listening passenger positions with additional amplification. The main component of an ICC system is an acoustic echo (or feedback) cancellation (AEC). It estimates the impulse response (IR) of the loudspeaker-enclosure-microphone (LEM) system in the car to calculate an estimated echo signal and subtract it from the microphone signal. By that, a widely echo-free speech component is obtained.

Most state-of-the-art ICC systems, as for example shown in [1], use a *single*-channel AEC, thereby being able to provide an easier communication inside the car cabin while increased driving noises are present. A different approach has been shown in [2] where another scenario is focused: passengers having a conversation while stereo music is played from the audio player or FM radio. To give sufficient consideration to this scenario, a *stereo*-channel AEC (SAEC) based on the frequency domain adaptive Kalman filter is modified in such a way that it is suitable for the low delay requirements of ICC systems. A great benefit of the used algorithm is the robust performance in stereo scenarios without additional decorrelation means [3].

In the closed acoustic loop of an ICC system, the performance of the AEC has a crucial impact. Increasing the AEC's ability to cancel the echoes (i.e., increasing the echo return loss enhancement (ERLE)), allows the system to become more robust and stable, and thus to reamplify the desired speech signals at a higher gain. One option to increase the ERLE is the additional use of a postfilter for residual echo suppression (RES) subsequent to the AEC. Based on the findings in [4] and [5], an RES postfilter for, e.g., hands-free telephony systems using the *multi*-channel Kalman filter, has been proposed in [6]. The approach exploits a tight relation between the RES coefficients and the stepsizes within the Kalman AEC, thereby allowing for an efficient way to obtain the RES postfilter coefficients. The proposed approach seems as well suitable for the use in a stereo ICC system, thereby improving the system's overall performance.

The remainder of this paper is structured as follows: First, the underlying signal model for a stereo-channel system is given in the time-domain. The formulations of two efficient residual echo postfilters are then given in the frequency domain. Subsequently, the performance of both postfilters is shown. Finally, the paper is concluded with a discussion of the results.

Signal Model

The postfilter we propose for the use in a stereo ICC system has been proposed for multi-channel hands-free telephony systems in [6] and is now outlined for the stereo-channel case. The underlying signal model for a stereo-channel system is as follows. The microphone signal is given as

$$\begin{aligned} y(n) &= s(n) + d_1(n) + d_2(n) \\ &= s(n) + \mathbf{h}_1^T \mathbf{x}_1(n) + \mathbf{h}_2^T \mathbf{x}_2(n), \end{aligned} \tag{1}$$

with near-end signal $s(n)$ and the two echo signals $d_j(n)$, $j \in \mathcal{I} = \{1, 2\}$. Echo signals $d_j(n)$ consist of the K latest loudspeaker samples

$$\mathbf{x}_j(n) = [x_j(n), \ldots, x_j(n-(K-1))]^T,\ j \in \mathcal{I}, \tag{2}$$

and the IR from loudspeaker to microphone (here written as time-invariant) $\mathbf{h}_j = [h_{j,0}, \ldots, h_{j,K-1}]^T, j \in \mathcal{I}$.

This model can be expanded to vectorial notation, then containing the K latest samples

$$\mathbf{y}(n) = \mathbf{s}(n) + \mathbf{X}_1^T(n)\mathbf{h}_1 + \mathbf{X}_2^T(n)\mathbf{h}_2 \tag{3}$$

with

$$\begin{aligned} \mathbf{y}(n) &= [y(n), \ldots, y(n-(K-1))]^T, \\ \mathbf{s}(n) &= [s(n), \ldots, s(n-(K-1))]^T, \\ \mathbf{X}_j(n) &= [\mathbf{x}_j(n), \ldots, \mathbf{x}_j(n-(K-1))],\ j \in \mathcal{I}. \end{aligned} \tag{4}$$

The convolution operations of SAEC and subsequent RES can be written similarly:

$$\hat{s}(n) = \big(\mathbf{y}^T(n) - \hat{\mathbf{h}}_1^T\mathbf{X}_1(n) - \hat{\mathbf{h}}_2^T\mathbf{X}_2(n)\big)\mathbf{g}. \qquad (5)$$

The equation describes the two steps to obtain a (widely) echo-free enhanced signal: the subtraction of estimated echo signals from the microphone signal using the estimated IRs $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$, $\hat{\mathbf{h}}_j = [\hat{h}_{j,0}, \ldots, \hat{h}_{j,K-1}]^T, j \in \mathcal{I}$, and the subsequent convolution with RES filter $\mathbf{g} = [g_0, \ldots, g_{K-1}]^T$.

Postfilter in the Frequency Domain

Based on the detailed time-domain derivation shown in [6] and in line with [7], the K-point Fourier transform with bin index k and frame index ℓ of the optimal RES filter is given as

$$\begin{aligned} G(\ell,k) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_1(\ell,k)\Phi_{1,2}(\ell,k)X_2^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{ss}(\ell,k). \end{aligned} \qquad (6)$$

$\Phi_{ss}(\ell,k)$ is the near-end signal's power spectral density, and $\Phi_{i,j}(\ell,k)$ are the so-called residual echo power transfer functions (cf. [4, p. 1143]).

To find an efficient formulation of the RES filter, we make use of the SAEC stepsizes for the frequency domain adaptive Kalman filter (cf. [8] and [6], with minor approximations) defined as

$$\begin{aligned} \mu_{i,j}(\ell,k) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_1(\ell,k)\Phi_{1,2}(\ell,k)X_2^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{i,j}(\ell,k). \end{aligned} \qquad (7)$$

A direct comparison of the equations (6) and (7) shows a tight relation: The optimal RES filter coefficients can be obtained very efficiently from the stepsizes as

$$G_{\mathrm{PF1}}(\ell,k) = 1 - \Big(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} X_i(\ell,k)\mu_{i,j}(\ell,k)X_j^*(\ell,k)\Big). \qquad (8)$$

A similar relation has as well been proposed by Enzner et al. for the *single*-channel case [4, eq. (77)], thus not including any cross-channel terms.

A second and differing postfilter formulation can be found when applying an additional assumption: The unpredictable residual parts of the two IRs $\mathbf{h}_1$ and $\mathbf{h}_2$ can be treated as statistically not correlated (see also [6]). In the frequency domain this correlation is described by the residual echo power transfer functions $\Phi_{1,2}(\ell,k)$ and $\Phi_{2,1}(\ell,k)$. As consequence equation (6) reduces to

$$\begin{aligned} G(\ell,k) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{ss}(\ell,k). \end{aligned} \qquad (9)$$

We deduce the efficient formulation of the optimal postfilter coefficients with zero-correlation assumption as

$$G_{\mathrm{PF2}}(\ell,k) = 1 - \Big(\sum_{j\in\mathcal{I}} X_j(\ell,k)\mu_j(\ell,k)X_j^*(\ell,k)\Big) \qquad (10)$$

with

$$\begin{aligned} \mu_j(\ell,k) = \big(&\Phi_{ss}(\ell,k) + X_1(\ell,k)\Phi_{1,1}(\ell,k)X_1^*(\ell,k) \\ &+ X_2(\ell,k)\Phi_{2,2}(\ell,k)X_2^*(\ell,k)\big)^{-1} \cdot \Phi_{j,j}(\ell,k) \end{aligned} \qquad (11)$$

being a separate stepsize definition to be used only for the RES postfilter.

Performance

The performance of the postfilters is shown in an experimental setup similar to an automotive stereo hands-free system. The 'far-end' signal is an FM radio speaker and the 'near-end' signal is a person on the driver position. This setup is comparable to an ICC system where an FM radio signal is present, the ICC processing is activated, but the computed enhanced signal is *not* additionally mixed to the FM Radio signal and played back into the car cabin. By doing so, we follow ITU-T Recommendation P.501 and additionally allow for a solid judgment of the algorithm and postfilter performance without having to consider effects caused by the ICC feedback loop.

The underlying SAEC is configured to a robust hands-free configuration [8]: sampling frequency 16 kHz, DFT size $K = 1024$, frame shift $R = 256$, forgetting factor $A = 0.998$, overestimation factor $\lambda = 1.5$ and $\mathbf{\Psi}_{ss}$ smoothing factor $\beta = 0.5$. Linear convolution in the echo path is ensured by using only $K - R$ coefficients to cover the estimated IRs of the echo path. Subsequently, the postfilter is applied to the SAEC's time-domain output in a separate windowing structure. Using square root Hann windows of length $2R$ with 50% overlap and zero-padding for a K-point DFT, the K postfilter coefficients are applied in the frequency domain. Moreover, the postfilter coefficients are subject to some practically motivated constraints to avoid artifacts: first order smoothing over time with factor 0.5 and gain limitation to the value range $[0.05, \ldots, 1]$.

The test signals from ITU-T Recommendation P.501 [9, Secs. 7.3.5, 7.3.7] are used to simulate a double-talk scenario as it can be seen in the waveforms at the bottom of Figure 1. The 'FM Radio Speaker' is extended to a stereo signal by convolving it with two randomly generated 'far-end' IRs. The Loudspeaker-Enclosure-Microphone is modeled by using two randomly generated IRs, as the former, with exponential energy decay and a reverberation time of $T_{60} = 50$ ms. The signals 'FM Radio Speaker' and 'Driver' are set to the same level and, additionally, the provided *in-car noise* is added as near-end noise at an SNR of 15 dB. To simulate a possible change in the LEM path, the impulse responses are challengingly switched after 31.5 s.

Beside the simulated speech signals, Figure 1 provides the system distance d_{sys} and overall performance in terms of echo return loss enhancement (ERLE). The SAEC algorithm without postfilter reaches up to 30 dB ERLE in single-talk, ca. 15 dB at far-end barge-ins, and

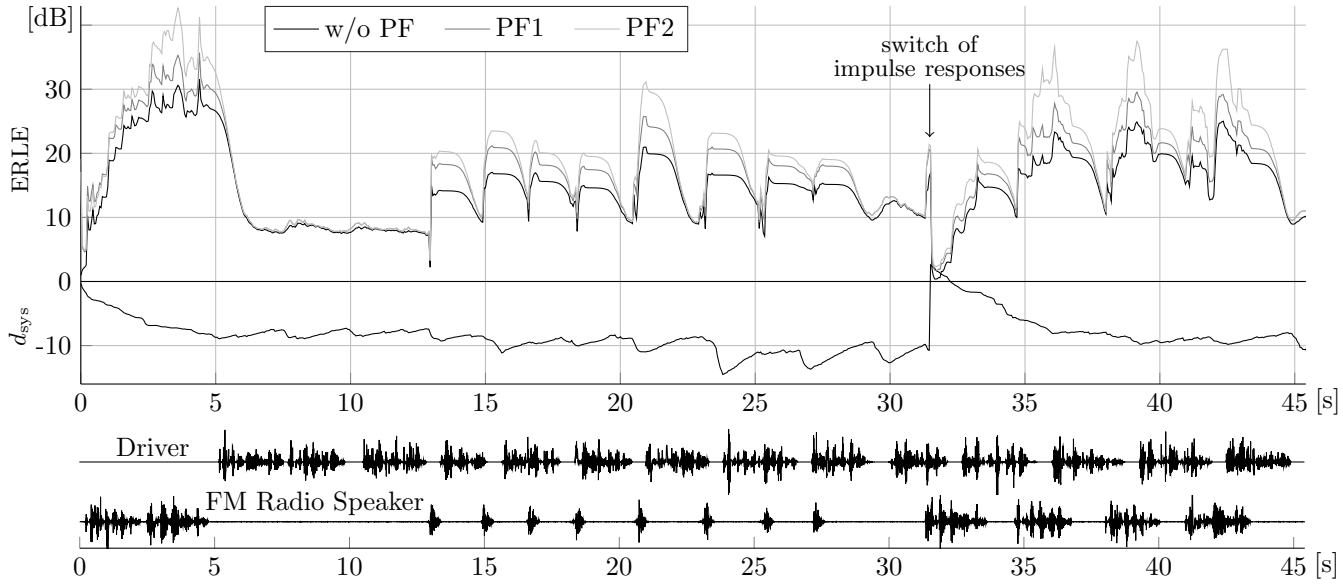

Figure 1: Performance without RES and with postfilters G_{PF1} and G_{PF2}: ERLE (top curves) and system distance (center curve). Speech waveforms (bottom) include single-talk (0-10 s), single-talk with barge-ins (10-30 s), double-talk (30-45 s).

around 20 dB during double-talk. Since the system distance does not depend on the postfilter, but only the IR estimation of the SAEC itself, it is the same for all three approaches. The fast convergence to a system distance value of about –10 dB can be seen in the center curve. In the simulated scenario the use of RES postfilter G_{PF1} increases the ERLE in all speech sections up to 5 dB. The use postfilter G_{PF2} leads to a further improvement during all sections: Values above 40 dB ERLE are reached during single-talk. During double-talk the postfilter G_{PF2} achieves up to 13 dB of additional echo suppression.

Discussion

In this paper, we proposed two efficient residual echo suppression approaches for a *stereo*-channel ICC system using the frequency-domain adaptive Kalman filter AEC. Though noting that the experiments are set in the context of an automotive hands-free system, the results reveal a clear improvement in terms of echo return loss enhancement. With about 5 dB additional RES for the first approach, and up to 13 dB of additional echo suppression for the second approach, both approaches show considerable potential to also increase the robustness of a stereo ICC system. For future research, the next interesting step could be the integration into a complete ICC system or a respective simulation environment. This should include the closed feedback loop as well as a postfilter windowing structure that considers the low delay requirements of in-car communication systems.

References

[1] P. Bulling, K. Linhard, A. Wolf, and G. Schmidt, "Stepsize Control for Acoustic Feedback Cancellation Based on the Detection of Reverberant Signal Periods and the Estimated System Distance," in *Proc. of INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 176–180.

[2] J. Franzen and T. Fingscheidt, "A Delay-Flexible Stereo Acoustic Echo Cancellation for DFT-Based In-Car Communication (ICC) Systems," in *Proc. of INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 181–185.

[3] S. Malik and G. Enzner, "Recursive Bayesian Control of Multichannel Acoustic Echo Cancellation," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 619–622, Nov. 2011.

[4] G. Enzner and P. Vary, "Frequency-Domain Adaptive Kalman Filter for Acoustic Echo Control in Hands-Free Telephones," *Signal Processing (Elsevier)*, vol. 86, no. 6, pp. 1140–1156, June 2006.

[5] F. Kuech, E. Mabande, and G. Enzner, "State-Space Architecture of the Partitioned-Block-Based Acoustic Echo Controller," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 1295–1299.

[6] J. Franzen and T. Fingscheidt, "An Efficient Residual Echo Suppression for Multi-Channel Acoustic Echo Cancellation Based on the Frequency-Domain Adaptive Kalman Filter," *accepted at ICASSP*, Calgary, Canada, Apr. 2018.

[7] S. Malik, *Bayesian Learning of Linear and Nonlinear Acoustic System Models in Hands-free Communication*, Ph.D. thesis, Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany, Oct. 2012.

[8] M. A. Jung, S. Elshamy, and T. Fingscheidt, "An Automotive Wideband Stereo Acoustic Echo Canceler using Frequency-Domain Adaptive Filtering," in *Proc. of EUSIPCO*, Lisbon, Portugal, Sept. 2014, pp. 1452–1456.

[9] "ITU-T Recommendation P.501, Test signals for use in telephonometry," ITU, Jan. 2012.