AN AUTOMOTIVE WIDEBAND STEREO ACOUSTIC ECHO CANCELER USING FREQUENCY-DOMAIN ADAPTIVE FILTERING

Marc-André Jung, Samy Elshamy, and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany
{m-a.jung,s.elshamy,t.fingscheidt}@tu-bs.de

ABSTRACT

We present an improved state-space frequency-domain acoustic echo canceler (AEC), which makes use of Kalman filtering theory to achieve very good convergence performance, particularly in double talk. Our contribution can be considered threefold: The proposed approach is designed to suit an automotive wideband overlap-save (OLS) setup, to operate best in this distinctive use case. Second, we provide a temporal smoothing and overestimation approach for two particular noise covariance matrices to improve echo return loss enhancement (ERLE) performance. Furthermore, we integrate an adapted perceptually transparent decorrelation preprocessor, which makes use of human insensitivity against appropriately chosen frequency-selective phase modulation, to improve robustness against far-end impulse response changes.

Index Terms— AEC, automotive, wideband, FDAF, decorrelation preprocessor

1. INTRODUCTION

Speech telecommunication has to cope with a number of possible sources for quality degradation, with disturbing acoustic echo as one of the prominent ones. To deal with that, acoustic echo canceler (AEC) technology is widely used. For this purpose oftentimes adaptive filters [1, 2] are employed to estimate an electric replica of the near-end echo path to reduce the disturbing echo component from the microphone signal by subtraction.

As hands-free telecommunication systems evolved to wideband HD voice services over the past years, new challenges came up, such as tougher quality requirements or higher computational complexity [3]. Whereas early systems commonly used time-domain algorithms of type normalized least mean squares (NLMS) [4] or similar, performance restrictions due to the higher bandwidth and poor convergence performance inspired block-based approaches [5–7] or convergence-optimized approaches [7], like recursive least squares (RLS) [8, 9] or affine projection (AP) [4, 10, 11] algorithms. To improve the ability to track changes of the impulse responses, oftentimes use of Kalman filter theory [7, 12] is made. Due to the fact, that parameters such as step size can be frequency-dependent, improvement in terms of performance, robustness, or convergence speed can be expected if at least the parameter adaptation is made in the frequency domain, hence resulting in so-called frequency-domain adaptive filtering (FDAF) algorithms [13].

If more than one channel is present, which might be the case when listening to FM radio while hands-free speech is acquired for automatic speech recognition or telephony, or because a multichannel telecommunication system is used (e.g., for teleconferencing), the AEC also has to operate in a plurality of channels. This is to estimate accurate replicas of the echo path signals in a unique way, which means, that the estimated filters match the corresponding real echo path impulse responses. In case the excitation signals are highly correlated, however, in general only the overall error energy is minimized, which does not necessarily imply a unique solution. Due to this, oftentimes decorrelation preprocessors are used to decorrelate the excitation signals, hence resulting in an improved misalignment score [14–16].

Since telecommunication services are of widespread importance in cars, automotive hands-free systems and therefore high-quality AEC algorithms are a demanded feature and even mandatory in many countries. To allow for conversations of high technical quality and low driver distraction, full-duplex capabilities are desirable [17]. This is only achievable, if the AEC algorithm provides robustness against near-end disturbances as they are not only present during double-talk periods [18, 19].

Our approach is based on [20], however, with several modifications and extensions being made. The algorithm has been adapted to an automotive stereo AEC approach with HD voice capabilities. Along with the state-space frequency-domain Kalman methodology, a perceptually optimized decorrelation preprocessor offers great robustness against changes of the far-end acoustics, very good tracking and convergence capabilities, and excellent double-talk performance. Temporal smoothing of the measurement noise covariance matrices and tuning of the process noise covariance matrices further increases echo return loss enhancement (ERLE) performance.

The organization of the paper is as follows: Section 2 describes a stereo acoustic echo canceler based on state-space frequency-domain adaptive filtering. A perceptually motivated decorrelation preprocessor is introduced to improve robustness against far-end acoustics changes. In Section 3 noise and speech simulation results for an automotive application are shown which put the proposed approach in relation to two baseline approaches. We then conclude our findings in Section 4.

2. STEREO FDAF

This section introduces the proposed algorithmic approach for stereo AEC and its underlying system model. References to two baseline approaches [20, 21] are made and a perceptually motivated decorrelation preprocessor is introduced.

2.1. System Model

In Fig. 1 the system model of a common stereo AEC system is depicted, with two loudspeakers and in this case one near-end microphone in the receiving room—modeled by two echo path impulse responses $h_1(n)$ and $h_2(n)$, with discrete sample index n—and two far-end microphones in the transmission room. The acoustic paths between far-end speaker and the two corresponding microphones are modeled by the impulse responses $h'_1(n)$ and $h'_2(n)$. At the near-end / far-end microphones a linear superposition of speech components ($s(n)$ / $s'_1(n)$, $s'_2(n)$), noise components ($n(n)$ / $n'_1(n)$, $n'_2(n)$), and echo components ($d_1(n)$, $d_2(n)$) is considered.

Two finite impulse response (FIR) filters take the loudspeaker signals ($x_1(n)$, $x_2(n)$) into account as reference to adapt their filter coefficients ($\hat{h}_1(n)$, $\hat{h}_2(n)$) by means of the error signal $e(n)$, which is a result of the difference between the near-end microphone signal $y(n) = \big(d_1(n) + d_2(n)\big) + s(n) + n(n)$ and the estimated echo signals $\hat{d}_j(n) = x_j(n) * \hat{h}_j(n)$, with channel index $j \in \{1, 2\}$, hats denoting estimated variables, and $*$ denoting a convolution.

We write vector and matrix entities as bold letters, scalars as normal letters, frequency-domain entities or constants as capital letters.

2.2. Algorithmic Approach

Notations and Initializations

For both channels the corresponding loudspeaker signal frame

$$\begin{aligned}\mathbf{x}_j(\ell) = \Big[&x_j\big((\ell-1)\cdot R\big),\ \ldots,\\ &x_j\big((\ell-1)\cdot R + K - R - 1\big),\\ &x_j\big((\ell-1)\cdot R + K - R\big),\\ &\ldots, x_j\big((\ell-1)\cdot R + K - 1\big)\Big]^T\end{aligned} \tag{1}$$

is initially composed as a vector of $K-R$ zeros followed by the first R samples of the speaker signals, with frame index $\ell \in \{1, 2, \ldots\}$, frame shift R, discrete Fourier transform (DFT) length K, and transpose operator T.

By applying the K-point DFT matrix $\mathbf{F}_{K\times K}$ and writing the result into a matrix main diagonal, the DFT-domain loudspeaker signal of channel j results in the matrix

$$\mathbf{X}_j(\ell) = \operatorname{diag}\big\{\mathbf{F}_{K\times K} \cdot \mathbf{x}_j(\ell)\big\}. \tag{2}$$

The $R \times 1$ near-end microphone signal $\mathbf{y}(\ell) = \big[y\big((\ell-1)\cdot R\big),\ \ldots,\ y\big((\ell-1)\cdot R + R - 1\big)\big]^T$, however, is first multiplied with the $K \times R$ overlap-save (OLS) projection matrix $\mathbf{Q} = (\mathbf{0}_{R\times K-R}\ \mathbf{I}_{R\times R})^T$, consisting of the zero matrix $\mathbf{0}$ and unity matrix $\mathbf{I}$, and then being transformed into the DFT domain by applying $\mathbf{F}_{K\times K}$, leading to the $K \times 1$ vector

$$\mathbf{Y}(\ell) = \mathbf{F}_{K\times K}\mathbf{Q} \cdot \mathbf{y}(\ell). \tag{3}$$

Furthermore, the first-order Markov model forgetting factors A_j, the DFT-domain AEC filter coefficients $\hat{\mathbf{H}}_j(\ell)$, state error covariance submatrices $\mathbf{P}_{j,i}(\ell)$, and the process noise covariance submatrices $\mathbf{\Psi}^{\Delta}_{j,i}(\ell)$ (cf. [20]) are initialized according to

$$\begin{aligned}&\left.\begin{aligned}A_j &= 0.998\\ \hat{\mathbf{H}}_j(\ell=0) &= \mathbf{0}_{K\times 1}\end{aligned}\right\} j \in \{1, 2\}\\ &\left.\begin{aligned}\mathbf{P}_{j,i}(\ell=0) &= \mathbf{I}_{K\times K}\\ \mathbf{\Psi}^{\Delta}_{j,i}(\ell=0) &= \mathbf{0}_{K\times K}\end{aligned}\right\} \forall j, i \in \{1, 2\}.\end{aligned} \tag{4}$$

Now a prediction step and a correction step are carried out in alternating fashion for all frames $\ell \in \{1, 2, \ldots\}$.

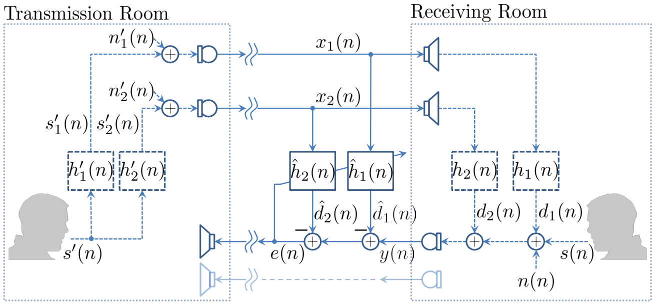

Fig. 1. System model of a stereo AEC system.

Prediction Step

The state of the AEC filter coefficients of channel j is predicted according to

$$\boxed{\hat{\mathbf{H}}^{+}_j(\ell) = A_j \hat{\mathbf{H}}_j(\ell-1) \qquad \forall j}\,, \tag{5}$$

with $(\cdot)^{+}$ denoting a predicted variable. The predicted state error covariances $\mathbf{P}^{+}_{j,i}$ are computed for the intra-channel ($j = i$) and cross-channel ($j \neq i$) case according to:

$$\begin{aligned}\mathbf{P}^{+}_{1,1}(\ell) &= A_1 A_1 \mathbf{P}_{1,1}(\ell-1) + \lambda \mathbf{\Psi}^{\Delta}_{1,1}(\ell-1)\\ \mathbf{P}^{+}_{1,2}(\ell) &= A_1 A_2 \mathbf{P}_{1,2}(\ell-1) + \lambda \mathbf{\Psi}^{\Delta}_{1,2}(\ell-1)\\ \mathbf{P}^{+}_{2,1}(\ell) &= A_2 A_1 \mathbf{P}_{2,1}(\ell-1) + \lambda \mathbf{\Psi}^{\Delta}_{2,1}(\ell-1)\\ \mathbf{P}^{+}_{2,2}(\ell) &= A_2 A_2 \mathbf{P}_{2,2}(\ell-1) + \lambda \mathbf{\Psi}^{\Delta}_{2,2}(\ell-1),\end{aligned} \tag{6}$$

with an overestimation factor λ.

Following [20], only the intra-channel process noise covariance submatrices $\mathbf{\Psi}^{\Delta}_{1,1}$ and $\mathbf{\Psi}^{\Delta}_{2,2}$ are updated:

$$\begin{aligned}\mathbf{\Psi}^{\Delta}_{1,1}(\ell-1) &= (1 - A_1^2)\big[\hat{\mathbf{H}}_1(\ell-1)\hat{\mathbf{H}}_1^H(\ell-1) + \mathbf{P}_{1,1}(\ell-1)\big]\\ \mathbf{\Psi}^{\Delta}_{2,2}(\ell-1) &= (1 - A_2^2)\big[\hat{\mathbf{H}}_2(\ell-1)\hat{\mathbf{H}}_2^H(\ell-1) + \mathbf{P}_{2,2}(\ell-1)\big]\end{aligned} \tag{7}$$

with $(\cdot)^H$ being the Hermitian transpose.

Correction Step

To incorporate the measurement at the current frame instance, the predicted filter coefficient states are corrected by the DFT-domain error signals, weighted by the corresponding Kalman gain diagonal matrices $\mathbf{K}_j(\ell)$

$$\boxed{\begin{aligned}\hat{\mathbf{H}}_1(\ell) = \hat{\mathbf{H}}^{+}_1(\ell) + \mathbf{K}_1(\ell)\cdot\Big[\mathbf{Y}(\ell) - \big(&\mathbf{G}\mathbf{X}_1(\ell)\hat{\mathbf{H}}^{+}_1(\ell)\\ &+ \mathbf{G}\mathbf{X}_2(\ell)\hat{\mathbf{H}}^{+}_2(\ell)\big)\Big]\\ \hat{\mathbf{H}}_2(\ell) = \hat{\mathbf{H}}^{+}_2(\ell) + \mathbf{K}_2(\ell)\cdot\Big[\mathbf{Y}(\ell) - \big(&\mathbf{G}\mathbf{X}_1(\ell)\hat{\mathbf{H}}^{+}_1(\ell)\\ &+ \mathbf{G}\mathbf{X}_2(\ell)\hat{\mathbf{H}}^{+}_2(\ell)\big)\Big]\end{aligned}} \tag{8}$$

with overlap-save constraint $\mathbf{G} = \mathbf{F}_{K\times K}\mathbf{Q}\mathbf{Q}^T\mathbf{F}^{-1}_{K\times K}$.

The Kalman gains are calculated as

$$\begin{aligned}\mathbf{K}_1(\ell) &= \boldsymbol{\mu}_{1,1}(\ell)\mathbf{X}_1^H(\ell) + \boldsymbol{\mu}_{1,2}(\ell)\mathbf{X}_2^H(\ell)\\ \mathbf{K}_2(\ell) &= \boldsymbol{\mu}_{2,1}(\ell)\mathbf{X}_1^H(\ell) + \boldsymbol{\mu}_{2,2}(\ell)\mathbf{X}_2^H(\ell)^{,}\end{aligned} \tag{9}$$

making use of the near-optimal step-size diagonal matrix $\boldsymbol{\mu}_{j,i}$:

$$\begin{aligned}\boldsymbol{\mu}_{1,1}(\ell) &= {}^{R}\!/\!{}_{K}\mathbf{P}^{+}_{1,1}(\ell)\mathbf{D}^{-1}(\ell)\\ \boldsymbol{\mu}_{1,2}(\ell) &= {}^{R}\!/\!{}_{K}\mathbf{P}^{+}_{1,2}(\ell)\mathbf{D}^{-1}(\ell)\\ \boldsymbol{\mu}_{2,1}(\ell) &= {}^{R}\!/\!{}_{K}\mathbf{P}^{+}_{2,1}(\ell)\mathbf{D}^{-1}(\ell)\\ \boldsymbol{\mu}_{2,2}(\ell) &= {}^{R}\!/\!{}_{K}\mathbf{P}^{+}_{2,2}(\ell)\mathbf{D}^{-1}(\ell)\end{aligned} \tag{10}$$

The diagonality of $\mathbf{P}^{+}_{j,i}(\ell)$ renders $\mathbf{D}(\ell)$ diagonal, too, and thus simplifies its inversion:

$$\begin{aligned}\mathbf{D}(\ell) = {}^{R}\!/\!{}_{K}\big[&\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,1}(\ell)\mathbf{X}_1^{H}(\ell)\\ &+\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,2}(\ell)\mathbf{X}_2^{H}(\ell)\\ &+\mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,1}(\ell)\mathbf{X}_1^{H}(\ell)\\ &+\mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,2}(\ell)\mathbf{X}_2^{H}(\ell)\big] + \boldsymbol{\Psi}^{S}(\ell)\end{aligned} \tag{11}$$

We calculate the measurement noise covariance matrix $\boldsymbol{\Psi}^{S}(\ell)$ in turn as

$$\begin{aligned}\boldsymbol{\Psi}^{S}(\ell) =&(1-\beta)\cdot\Big(\tilde{\mathbf{E}}(\ell)\tilde{\mathbf{E}}^{H}(\ell)\\ &+{}^{R}\!/\!{}_{K}\big[\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,1}(\ell)\mathbf{X}_1^{H}(\ell)\\ &\quad+\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,2}(\ell)\mathbf{X}_2^{H}(\ell)\\ &\quad+\mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,1}(\ell)\mathbf{X}_1^{H}(\ell)\\ &\quad+\mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,2}(\ell)\mathbf{X}_2^{H}(\ell)\big]\Big) + \beta\cdot\boldsymbol{\Psi}^{S}(\ell-1)\end{aligned} \tag{12}$$

by means of smoothing over time, with empirical smoothing constant $\beta = 0.5$, and the *preliminary* error vector

$$\tilde{\mathbf{E}}(\ell) = \mathbf{Y}(\ell) - \big[\mathbf{G}\mathbf{X}_1(\ell)\hat{\mathbf{H}}_1^{+}(\ell) + \mathbf{G}\mathbf{X}_2(\ell)\hat{\mathbf{H}}_2^{+}(\ell)\big]. \tag{13}$$

At this point the recursion has come to an end and (8) is completely depicted. By the help of (6) and (9) the predicted state error covariance matrices can be corrected to

$$\begin{aligned}\mathbf{P}_{1,1}(\ell) &= \mathbf{P}^{+}_{1,1}(\ell) - {}^{R}\!/\!{}_{K}\mathbf{K}_1(\ell)\big[\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,1}(\ell) + \mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,1}(\ell)\big]\\ \mathbf{P}_{1,2}(\ell) &= \mathbf{P}^{+}_{1,2}(\ell) - {}^{R}\!/\!{}_{K}\mathbf{K}_1(\ell)\big[\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,2}(\ell) + \mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,2}(\ell)\big]\\ \mathbf{P}_{2,1}(\ell) &= \mathbf{P}^{+}_{2,1}(\ell) - {}^{R}\!/\!{}_{K}\mathbf{K}_2(\ell)\big[\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,1}(\ell) + \mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,1}(\ell)\big]\\ \mathbf{P}_{2,2}(\ell) &= \mathbf{P}^{+}_{2,2}(\ell) - {}^{R}\!/\!{}_{K}\mathbf{K}_2(\ell)\big[\mathbf{X}_1(\ell)\mathbf{P}^{+}_{1,2}(\ell) + \mathbf{X}_2(\ell)\mathbf{P}^{+}_{2,2}(\ell)\big].\end{aligned} \tag{14}$$

Finally the error (i.e., enhanced) signal can be determined as follows

$$\mathbf{E}(\ell) = \mathbf{Y}(\ell) - \big[\mathbf{G}\mathbf{X}_1(\ell)\hat{\mathbf{H}}_1(\ell) + \mathbf{G}\mathbf{X}_2(\ell)\hat{\mathbf{H}}_2(\ell)\big]. \tag{15}$$

2.3. Decorrelation Preprocessor

As already known from literature [14, 15], a *non-uniqueness problem* exists for multi-channel AEC, which leads to sub-optimal filter adaptation if the loudspeaker signals are highly correlated. In such a case, the filters converge to a state which minimizes the energy of the common error signal $e(n)$, not necessarily identifying the echo path impulse responses $h_j(n)$. Therefore, a change of the transmission room impulse responses will lead to a distinct drop of convergence. Different preprocessing schemes exist, which try to optimize the trade-off between convergence enhancement, subjective sound quality, and complexity, by using addition of uncorrelated signals, decorrelation filters, frequency shifting, or comb filters [14, 16].

We have chosen a perceptually motivated decorrelation preprocessor which uses phase modulation of opposite direction to decorrelate the two loudspeaker signals $x_j(n), j \in \{1, 2\}$, as proposed in [16]. In contrast to, e.g., a classical nonlinear filter, artifacts are far less audible. A modulation frequency of $f_m = 1\,\text{Hz}$ was used, with a frequency-dependent modulation amplitude of $\pm 10^\circ$ below $1\,\text{kHz}$, a linear increase up to $\pm 40^\circ$ at $2\,\text{kHz}$, a further linear increase up to $\pm 90^\circ$ at $2.5\,\text{kHz}$, and a constant $\pm 90^\circ$ above $2.5\,\text{kHz}$. This is done to reduce the perceived signal distortion whilst still decorrelating as much as possible where human perception is mostly insensitive to phase changes.

2.4. Distinction Against Baseline Approaches

The proposed approach is derived from the submatrix-diagonal multichannel state-space frequency-domain adaptive filter (SD-MCSSFDAF) [20] and its variationally-diagonalized version with implicit omission of cross-channel terms VD-MCSSFDAF [21], whereas in our case the intended use case is an automotive setup and additional modifications have been made.

Unlike in [20, 21], a perceptually motivated decorrelation approach described in [16] has been modified so that it can be used in our DFT-processing environment. Therefore, flat-top Hann windowing has been used to keep artifacts at a minimum. To improve the convergence speed of the proposed approach, an empirical overestimation factor of $\lambda = 1.5$ for $\boldsymbol{\Psi}^{\Delta}(\ell)$ was introduced (cf. (6)). To also improve the ERLE in the converged state, the parameter $\boldsymbol{\Psi}^{S}(\ell)$ is recursively smoothed over time in (12) using $\beta = 0.5$. Furthermore, the filter coefficients in (8) are constraint to length N to avoid circular artifacts.

In the calculation of $\boldsymbol{\Psi}^{S}(\ell)$, as it is shown in [21, eq. (46)] or in [22, eq. (23)], the state error covariances $\mathbf{P}_{j,i}(\ell)$ are needed. As these are not yet available, for our implementation of [20] and [21] the predicted terms $\mathbf{P}^{+}_{j,i}(\ell)$ are used instead (as in (12)).

3. PERFORMANCE EVALUATION

To grade the performance of the proposed approach in contrast to the baseline approaches, simulations have been carried out, adopting parameters from an automotive setup.

3.1. Automotive Simulation Setup

Simulations for an automotive use case have been carried out. In order to provide best reproducibility of our results, we decided to randomly generate all impulse responses and to equip them with exponential energy decay, so that a reverberation time of $T_{60} = 50\,\text{ms}$ has been achieved. In the transmission room a common white noise or speech signal is being convolved with the far-end impulse responses $h_1'(n)$ and $h_2'(n)$ to yield the far-end speech components $s_1'(n)$ and $s_2'(n)$ with an amplitude of $-26\,\text{dBov}$ each at the microphones. After adding *un*correlated white Gaussian sensor noise of amplitude $-66\,\text{dBov}$, the loudspeaker signals $x_j(n) = \big(s'(n) * h_j'(n)\big) + n_j'(n), j \in \{1, 2\}$, are available as AEC filter reference signals and receiving room excitation signals. To provide comparability amongst all evaluated approaches, a decorrelation of $x_j(n)$ takes place as soon as the far-end microphone signals are available. The excitation signals are convolved with the normalized echo path impulse responses $h_j(n)$ individually, to achieve the echo components $d_j(n)$. Together with the near-end speech component $s(n)$ with amplitude $-26\,\text{dBov}$ and the near-end car noise component $n(n)$ with

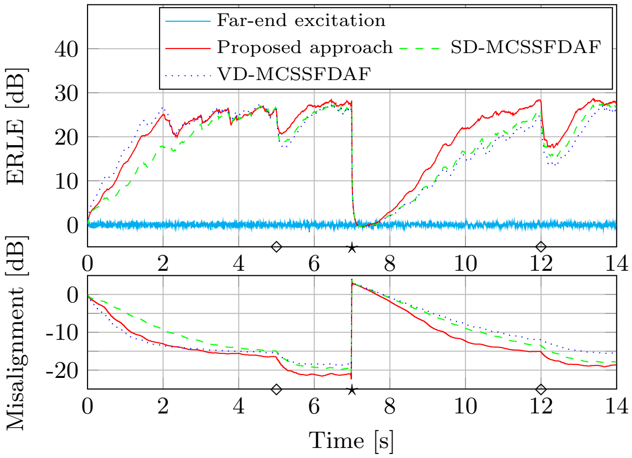

Fig. 2. Convergence behavior during white Gaussian noise excitation at -26 dBov per channel and near-end car noise at -41 dBov. Far-end impulse response switches (transmission room) indicated by ⋄, near-end impulse response switches indicated by ⋆.

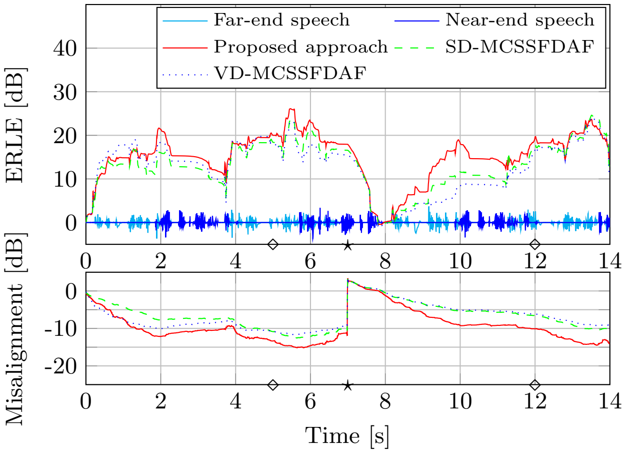

Fig. 3. Alternating / overlapping speech periods on the far and near end, each with amplitude -26 dBov (per channel). Near-end car noise at -41 dBov. Far-end impulse response switches (transmission room) indicated by ⋄, near-end impulse response switches indicated by ⋆.

amplitude -41 dBov the microphone signal $y(n)$ is subject to echo cancellation.

To examine both optimal and realistic scenarios a far-end single talk setup with white noise excitation (Fig. 2) and a speech setup with single and double talk periods (Fig. 3) have been chosen. Both setups include near-end car noise with amplitude -41 dBov. To represent changing acoustic conditions of the transmission and receiving room, the far-end impulse responses are switched to newly generated ones at times denoted by ⋄, whereas the time of a near-end impulse response switch is denoted by ⋆.

The parametrization chosen for all three approaches was: Forgetting factors $A_j = 0.998$, constants $\lambda = 1.5$ and $\beta = 0.5$, frame shift $R = 256$, DFT length $K = 1024$, and sample frequency $f_s = 16\,\text{kHz}$. In so doing, a maximum AEC filter impulse response length of $N = K - R = 768$ samples, corresponding to 48 ms, is achievable. The length of the impulse responses $h_j(n)$ and $h'_j(n)$ equals their reverberation time of $T_{60} = 50\,\text{ms}$.

3.2. Discussion

In Fig. 2 the convergence behavior of the proposed approach is compared with the two baseline approaches [20, 21] described in Section 2.4. It can be seen, that initial convergence[1] is at a very good level of around 1–1.5 s for both the proposed approach and VD-MCSSFDAF. The SD-MCSSFDAF, however, takes about 2 s to converge. All three approaches saturate to about 29 dB ERLE. The first far-end impulse response change (symbol ⋄ on the time axis) reveals the non-uniqueness problem, since all three approaches decline in ERLE, whereas the proposed approach shows more robustness compared to the baselines. This is also apparent from reconvergence speed with only the proposed approach being able to reconverge to a full extent two seconds after the first switch. The following near-end impulse response switch (symbol ⋆) again leads to a decline in ERLE, to now 0 dB for all approaches, from which the proposed approach is able to reconverge[1] in 2.5 s and the baseline approaches in about 3.5 s. The second far-end impulse response change (symbol ⋄ again) shows characteristics for the three approaches as before, whereas the VD-MCSSFDAF reacts even somewhat worse than before.

Fig. 3 shows the performance of all three approaches during a realistic conversation with speech at both ends, having some short periods of double talk. In this setup the good convergence and double-talk performance of the proposed approach is confirmed, only being outperformed by the VD-MCSSFDAF approach during the first second of initial convergence.

In the normalized misalignment[2] plots of Fig. 2 and Fig. 3 the effects of the proposed decorrelation approach of Section 2.3 can be clearly seen. The proposed approach outperforms both baselines by up to 5 dB in terms of misalignment, leading to higher robustness against far and near end impulse response changes.

4. CONCLUSIONS

We have proposed an automotive wideband stereo AEC algorithm, based on state-space frequency-domain adaptive Kalman filtering [20, 21], which excels in terms of performance and speech quality specifically during double talk, so that complete full-duplex capabilities are given. A good initial convergence speed of about 1.5 s and of about 2.5 s for the reconvergence case could be achieved even during double talk, so that reconvergence times could be reduced in comparison to the baseline approaches. Whereas computational complexity is higher compared to the variationally-diagonalized MCSSFDAF [21], it is on par with the SD-MCSSFDAF algorithm [20] in $\mathcal{O}$ notation.

Very high robustness against far-end impulse response changes is achieved by making use of a decorrelation preprocessor, which is able to effectively encounter the non-uniqueness problem without introducing perceptual disturbances to the loudspeaker signals: An improvement of up to 5 dB in terms of misalignment could be achieved, which in turn offers a greater robustness against the effects of a moving far-end speaker. The shown convergence and double-talk performance of the proposed approach can be considered very good for this automotive setting with high echo coupling and near-end car noise.

[1] The time needed to reach an ERLE of 20 dB. In this paper, ERLE is computed as shown in [20].

[2] We compute the normalized total misalignment of both channels according to $D(n) = 10\log\left[\sum_{j=1}^{2}||\mathbf{h}_j - \hat{\mathbf{h}}_j||^2 / \sum_{j=1}^{2}||\mathbf{h}_j||^2\right]$.

REFERENCES

[1] B. Widrow and P. N. Stearns, *Adaptive Signal Processing*, 1st ed. Englewood Cliffs, NJ: Prentice Hall, Mar. 1985.

[2] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice Hall International, Sep. 2002.

[3] C. Beaugeant, M. Schönle, and I. Varga, "Challenges of 16 kHz in Acoustic Pre- and Post-Processing for Terminals," *IEEE Communications Magazine*, vol. 44, no. 5, pp. 98–104, 2006.

[4] H.-C. Shin, A. Sayed, and W.-J. Song, "Variable Step-Size NLMS and Affine Projection Algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132–135, 2004.

[5] J. Lee and C. Un, "Block Realization of Multirate Adaptive Digital Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 105–117, 1986.

[6] J. J. Shynk, "Frequency-Domain and Multirate Adaptive Filtering," *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[7] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Hoboken, NJ: John Wiley & Sons, 2004.

[8] G. Carayannis, D. Manolakis, and N. Kalouptsidis, "A Fast Sequential Algorithm for Least-Squares Filtering and Prediction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, no. 6, pp. 1394–1402, 1983.

[9] J. Cioffi and T. Kailath, "Fast, Recursive-Least-Squares Transversal Filters for Adaptive Filtering," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 304–337, 1984.

[10] K. Ozeki and T. Umeda, "An Adaptive Filtering Algorithm Using an Orthogonal Projection to an Affine Subspace and its Properties," *Electron. Commun. Jpn.*, vol. 67, no. 5, pp. 19–27, 1984.

[11] S. Gay and S. Travathia, "The Fast Affine Projection Algorithm," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995)*, vol. 3, 1995, pp. 3023–3027.

[12] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Trans. ASME, J. Basic Eng.*, vol. 82, pp. 35–45, 1960.

[13] J. Shynk and R. Gooch, "Frequency-Domain Adaptive Pole-Zero Filtering," *Proceedings of the IEEE*, vol. 73, no. 10, pp. 1526–1528, Oct. 1985.

[14] M. Sondhi, D. Morgan, and J. Hall, "Stereophonic Acoustic Echo Cancellation – An Overview of the Fundamental Problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, 1995.

[15] J. Benesty, D. R. Morgan, and M. Mohan, "A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.

[16] J. Herre, H. Buchner, and W. Kellermann, "Acoustic Echo Cancellation for Surround Sound Using Perceptually Motivated Convergence Enhancement," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, vol. 1, 2007, pp. I–17–I–20.

[17] ITU-T, *Rec. P.1110: Wideband Hands-free Communication in Motor Vehicles*, International Telecommunication Union, Dec. 2009.

[18] E. Hänsler and G. Schmidt, Eds., *Speech and Audio Processing in Adverse Environments*. Berlin / Heidelberg, Germany: Springer, 2008.

[19] M.-A. Jung and T. Fingscheidt, *Smart Mobile In-Vehicle Systems: Next Generation Advancements*. New York: Springer Science+Business Media, 2014, ch. 6: A Wideband Automotive Hands-free System for Mobile HD Voice Services.

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

[21] S. Malik and J. Benesty, "Variationally Diagonalized Multichannel State-Space Frequency-Domain Adaptive Filtering for Acoustic Echo Cancellation," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, May 2013, pp. 595–599.

[22] S. Malik and G. Enzner, "Online Maximum-Likelihood Learning of Time-Varying Dynamical Models in Block-Frequency Domain," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, Mar. 2010, pp. 3822–3825.