EFFICIENT NONLINEAR ACOUSTIC ECHO CANCELLATION BY DUAL-STAGE MULTI-CHANNEL KALMAN FILTERING

Matthias Schrammen[1], *Stefan Kühl*[1], *Shmulik Markovich-Golan*[2], *Peter Jax*[1]

[1]Institute of Communication Systems (IKS), RWTH Aachen University, Germany
{*schrammen, kuehl, jax*}*@iks.rwth-aachen.de*
[2]Intel Communication and Devices Group, Intel corporation, Israel
shmulik.markovich-golan@intel.com

ABSTRACT

Mobile devices for hands-free speech communication often show significant nonlinear distortion in the sound emitted by their loudspeakers. Therefore, conventional linear echo cancellation is not sufficient for maintaining a high conversation quality. In this work we propose a nonlinear echo canceller that uses two serially cascaded adaptive filters to compensate for the nonlinear and linear echo. We show that a stable operation of the cascaded structure is achieved by using the multi-channel Kalman algorithm in the frequency domain with filtered-x references. By modelling a nonlinearity with memory we further improve the performance. To reduce the computational complexity of the proposed solution we derive an efficient decimation scheme by exploiting useful properties of the cascaded approach.

Index Terms— nonlinear acoustic echo cancellation, multi-channel Kalman filter, cascaded adaptive filters

1. INTRODUCTION

Modern mobile devices for speech communication are often operated in hands-free mode. This requires the loudspeaker to emit high sound pressure levels because the near-end listener is probably far away from the device and/or is in a noisy environment. At the same time, small form factors, low energy consumption, and cheap components for the loudspeaker and amplifier are requested. These circumstances often lead to significant nonlinear components in the sound emitted by the loudspeaker. Therefore, conventional acoustic echo cancellation (AEC) based on linear filters only is not sufficient and solutions for nonlinear AEC (NAEC) are needed to guarantee a sufficient quality of the conversation. Most solutions found in literature focus on generating artificial nonlinear versions of the far-end signal. However, there are also solutions that rely on measured voltage or current signals to acquire a nonlinear reference close to the real loudspeaker output [1]. The solutions generating nonlinear versions of the far-end signal can roughly be categorized into two approaches. The first one models the whole echo path as one nonlinear filter with eventually multiple parallel branches [2–4]. The second approach uses a serial cascade of a nonlinear and linear filter [5–7]. On one hand the adaptation control for the one-filter approach is easier than for the cascaded one. On the other hand the number of coefficients becomes very large, because the memory of the room acoustic path must be taken into account. This in turn often results in a slow convergence or in an undermodelled echo path, because the nonlinear order must be chosen small to guarantee sufficient convergence speed. For the cascaded structures the degrees of freedom (nonlinear memory and order) can be tailored more to the physical echo path, resulting in much less coefficients. However, the adaptation of the cascaded filters needs special treatment to guarantee stability and convergence. In [8] a cascaded model consisting of a power series with short memory for the nonlinear part and a longer linear filter for the room acoustic transmission was proposed. However, the adaptation was done in the time domain with the NLMS algorithm that needs a sophisticated stepsize control and initialization to be usable in practice. The problem of stepsize control in cascaded filter structures was solved in [9] by using two Kalman filters working in the time domain. One Kalman filter was used to adapt a memoryless preprocessor and the other one was used for the adaptation of the linear filter modelling the room acoustics. In Sec. 2 of this contribution we extend the model in [9] to a nonlinearity with memory as proposed in [8], describe the parallel approach for the adaptation and briefly discuss its properties. After that, in Sec. 3, we show how to improve the performance by using a new structure in the block-based frequency domain for the adaptation of the cascaded approach and reduce its complexity. The evaluation on measured signals is presented in Sec. 4.

This work was funded by Intel Mobile Communications.

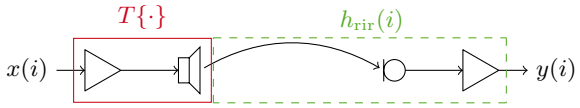


Fig. 1. Digital system model of the true echo path. The dashed green and solid red boxes indicate linear and nonlinear subsystems, respectively.

2. SIGNAL MODEL

Fig. 1 shows an all-digital block diagram of the true echo path. The digital far-end signal $x(i)$ is amplified and fed into the loudspeaker. The whole system containing amplifier and loudspeaker is modeled as one system $T\{\cdot\}$, which can be nonlinear and can contain memory. The acoustic transmission of the sound through the room, the microphone, and the amplifier is modeled as one linear FIR filter $h_{\mathrm{rir}}(i)$. The transmission relating the digital far-end signal $x(i)$ to the digital microphone signal $y(i)$ is modeled as

$$y(i) = h_{\mathrm{rir}}(i) * T\{x(i)\}, \tag{1}$$

where i is the discrete time index (see Fig. 1). For the nonlinear function $T\{\cdot\}$, describing the amplifier and the loudspeaker, an odd order power series of order P with memory is chosen according to

$$T\{x(i)\} = \sum_{l=0}^{\lfloor P/2 \rfloor} w_{p(l)}(i) * x^{p(l)}(i), \ \text{with}\, p(l) = 2l+1, \tag{2}$$

where $w_p(i)$ models the nonlinear memory of the p-th order. In reality a digital-to-analog (DA) and analog-to-digital (AD) conversion is also part of the system. We assume that both are linear and that the DA conversion is modelled by $w_p(i)$ as suggested in [2]. The AD conversion is easily incorporated in $h_{\mathrm{rir}}(i)$. Often $w_p(i)$ is modelled as a scalar only [3, 6, 9]. However, our experiments showed that it can be beneficial for the NAEC performance to allow for some nonlinear memory. The odd order power series was chosen because it achieved best results when compared to other expansions with the same maximum order P. However, the approach presented in this paper is not limited to the odd order power series, i.e., other bases like Legendre or Fourier expansions and even orders can be used as well. For mitigating the aliasing that occurs when calculating the powers $x^p(i)$ in the digital domain, we use an intermediate oversampling.

Conventionally, to circumvent difficulties arising with the adaptation of cascaded filters, the single-channel nonlinear system is often converted to a multi-channel linear system with nonlinear input signals. In the following we briefly introduce the full multi-channel Kalman (Full-MCK) filter proposed in [3], that uses this conversion and will serve as a state-of-the-art anchor for our proposed solution described in Sec. 3. The conversion is done by inserting (2) into (1) and moving $h_{\mathrm{rir}}(i)$ into the sum. Then, the effective impulse response $h_p(i) = h_{\mathrm{rir}}(i) * w_p(i)$ of the p-th order can be defined, where we assume that $h_{\mathrm{rir}}(i)$ and $w_p(i)$ are time-invariant during one frame to be able to realize the convolution. Now the Full-MCK in the frequency domain can be applied to adapt $h_p(i)$ with $x^p(i)$ as reference for the p-th channel [10]. Because the nonlinear references are not mutually orthogonal, we use the sub-diagonalized version of the Full-MCK as proposed in [3].

The multi-channel interpretation of the nonlinear echo path enables a convenient application of well-known algorithms for multi-channel acoustic echo cancellation. However, it over-models the physical system in terms of memory, because $w_p(i)$ is typically very short and in most cases even modelled as a scalar only. Then, the impulse responses $h_p(i)$ can simply be scaled versions of the same room impulse response $h_{\mathrm{rir}}(i)$. If N_P channels and an FFT-size of M are used, the Full-MCK has $M \cdot N_P$ degrees of freedom instead of $M + N_P$ as in the assumed model, where $N_P = \lfloor P/2 \rfloor + 1$ for the odd order power series. Second, by increasing the number of channels, the total number of coefficients that have to be adapted increases. This raises the complexity and slows down the convergence speed of the Full-MCK, which makes the tracking of time-variant acoustic scenarios more difficult. To improve the performance, we first introduce a structure for the adaptation of the cascaded model with nonlinear memory in Sec. 3 and then reduce its complexity in Sec. 3.1.

3. DUAL-STAGE MULTI-CHANNEL KALMAN FILTER

The proposed structure for NAEC shown in Fig. 2 maintains the cascaded structure of the true echo path model in the compensation path. Only the loudspeaker and amplifier are modeled as a multi-channel system with nonlinear references and memory $\hat{\boldsymbol{w}}_p(k)$ in stage 1 (S1), whereas the transmission after the loudspeaker is modeled as a single-channel system $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$ in stage 2 (S2). As two adaptive filters are concatenated serially we have to carefully choose the reference signals and the error signal provided to S1. For the operation of serially cascaded multi-channel nonlinear adaptive filters it is known from literature that both, the initialization and the stepsize control, must be treated with special care [5, 6, 11]. The proposed dual-stage solution uses the Kalman algorithm to adapt both, the nonlinear S1 and the linear S2. Therefore, this approach will be termed dual-stage multi-channel Kalman (DualStage-MCK) filter. The problem of stepsize control is alleviated because the Kalman algorithm already provides

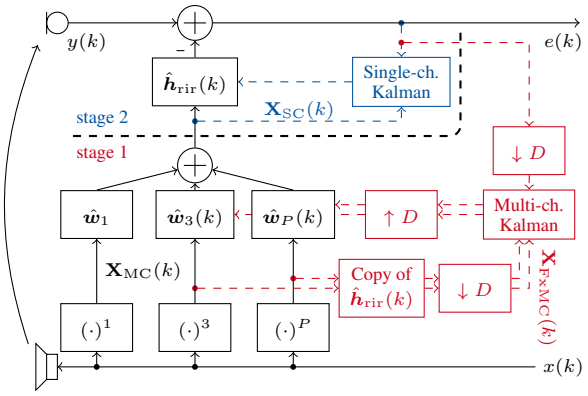


Fig. 2. Block diagram of the proposed dual-stage multi-channel Kalman (DualStage-MCK). The nonlinear stage 1 (S1) is adapted with a multi-channel Kalman filter using a Filtered-x nonlinear reference. The linear stage 2 (S2) is adapted with a single-channel Kalman filter.

a near-optimum stepsize control [3]. With ($\downarrow D$) and ($\uparrow D$) a decimation and interpolation is indicated in Fig. 2, respectively. This is necessary for the complexity reduction discussed in Sec. 3.1. However, in this section no decimation or interpolation will be applied yet, i.e. $D = 1$. For both stages we use the same overlap-save framework with frame shift R and short-time Fourier transform of size M as for the Full-MCK [3, 10, 12]. Therefore, we define the frame-wise reference signal of the p-th order, its Fourier transform and the multi-channel frequency-domain reference signal as

$$\boldsymbol{x}_{p,M}(k) = (x^p(kR - M + 1), x^p(kR - M + 2), \ldots, x^p(kR))^T$$

$$\mathbf{X}_p(k) = \mathrm{diag}\{\mathbf{F}_M \cdot \boldsymbol{x}_{p,M}(k)\} \in \mathbb{C}^{M \times M} \quad (3)$$

$$\mathbf{X}_{\mathrm{MC}}(k) = [\mathbf{X}_1(k), \mathbf{X}_3(k), \ldots, \mathbf{X}_P(k)] \in \mathbb{C}^{M \times MN_P}, \quad (4)$$

where k is the frame index, $\mathbf{F}_M$ and $\mathbf{F}_M^{-1}$ are the Fourier matrix of size $M \times M$ and its inverse, respectively. Now we start with the description of S2 and define the frequency domain representation of the filter coefficients to be estimated in S2 as

$$\hat{\boldsymbol{h}}_{\mathrm{rir}}(k) = \left(\hat{h}_{\mathrm{rir},1}(k), \hat{h}_{\mathrm{rir},2}(k), \cdots, \hat{h}_{\mathrm{rir},N_{\mathrm{lin}}}(k)\right)^T \quad (5)$$

$$\hat{\mathbf{H}}_{\mathrm{SC}}(k) = \mathbf{F}_M \begin{pmatrix} \hat{\boldsymbol{h}}_{\mathrm{rir}}(k) \\ \mathbf{0}_{M - N_{\mathrm{lin}}} \end{pmatrix} \in \mathbb{C}^{M \times 1}. \quad (6)$$

The constraining in (6) is necessary to avoid cyclic artifacts. The effective impulse response of both stages is of length $N_{\mathrm{nl}} + N_{\mathrm{lin}} - 1$, because it results from the convolution of $\hat{\boldsymbol{w}}_p(k)$ of length N_{nl} with $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$ of length N_{lin}. It follows that the maximum length of $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$ must be restricted to $N_{\mathrm{lin}} = M - R - N_{\mathrm{nl}} + 2$. With this we need the last R samples of microphone signal

$$\boldsymbol{y}_R(k) = (y(kR - R + 1), y(kR - R + 2), \ldots, y(kR))^T \quad (7)$$

to calculate the error signal and the reference signal of S2 as

$$\mathbf{E}(k) = \mathbf{F}_M \mathbf{Q}_R \, \boldsymbol{y}_R(k) - \mathbf{G}_R \mathbf{X}_{\mathrm{SC}}(k) \hat{\mathbf{H}}_{\mathrm{SC}}(k) \in \mathbb{C}^{M \times 1} \quad (8)$$

$$\mathbf{X}_{\mathrm{SC}}(k) = \mathrm{diag}\left\{\mathbf{G}_{\mathrm{nl}} \mathbf{X}_{\mathrm{MC}}(k) \hat{\mathbf{W}}'_{\mathrm{MC}}(k)\right\} \in \mathbb{C}^{M \times M}. \quad (9)$$

To avoid artifacts due to cyclic convolution the constraining matrices $\mathbf{G}_\chi$ are used, where $\mathbf{G}_\chi = \mathbf{F}_M \mathbf{Q}_\chi \mathbf{Q}_\chi^H \mathbf{F}_M^{-1}$ with $\chi \in \{R, \mathrm{nl}\}$ and the zero-padding matrices $\mathbf{Q}_{\mathrm{nl}} = (\mathbf{0}_{N_{\mathrm{nl}}-1} \quad \mathbf{I}_{M-N_{\mathrm{nl}}+1})^T$ and $\mathbf{Q}_R = (\mathbf{0}_{M-R} \quad \mathbf{I}_R)^T$. $\mathbf{I}_n$ is the identity matrix of size $n \times n$ and $\mathbf{0}_{M-n}$ is the zero-matrix of size $n \times (M - n)$. For the adaptation

of $\hat{\mathbf{H}}_{\mathrm{SC}}(k)$ the fully diagonalized single-channel Kalman algorithm described in [13] is used with a forgetting factor of 0.9999.

Now we describe the adaptation in S1. The filters $\hat{\boldsymbol{w}}_p(k)$ modelling the nonlinear memory $w_p(i)$ of the p-th order are adapted by a multi-channel Kalman filter for $p > 1$, i.e., the Kalman filter has $\tilde{N}_P = N_P - 1$ channels. The linear branch of S1 is excluded from the adaptation to prevent any ambiguity of the coefficients of $\hat{\boldsymbol{w}}_1$ and $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$, because both model a linear behavior. Therefore, $\hat{\boldsymbol{w}}_1$ is set to a constant delay of $i_0 = \mathrm{Round}[(N_{\mathrm{nl}} - 1)/2] + 1$ samples. This choice of $\hat{\boldsymbol{w}}_1$ ensures some degree of freedom for the nonlinear memory of the loudspeaker. All other filter coefficients are initialized with zeros. Generally, the nonlinear reference signals originating from the odd order power series expansion are correlated. To mitigate a negative impact on the convergence of the adaptive algorithm, we use the sub-diagonalized version of the multi-channel Kalman algorithm and a short length for the filters in the nonlinear stage. Thereby, a change of the expansion series and the use of an expensive adaptive orthogonalization scheme as suggested in [2] is avoided. The Kalman filter of S1 gets the same error signal $\mathbf{E}(k)$ as S2. We define the following frequency domain representation of the filter coefficients to be estimated in S1:

$$\hat{\boldsymbol{w}}_p(k) = (\hat{w}_{p,1}(k), \hat{w}_{p,2}(k), \ldots, \hat{w}_{p,N_{\mathrm{nl}}}(k))^T \tag{10}$$

$$\hat{\mathbf{W}}_p(k) = \mathbf{F}_M \begin{pmatrix} \hat{\boldsymbol{w}}_p(k) \\ \mathbf{0}_{M-N_{\mathrm{nl}}} \end{pmatrix} \in \mathbb{C}^{M\times 1} \tag{11}$$

$$\hat{\mathbf{W}}_{\mathrm{MC}}(k) = \left[\hat{\mathbf{W}}_3^T(k), \hat{\mathbf{W}}_5^T(k), \ldots, \hat{\mathbf{W}}_P^T(k)\right]^T \in \mathbb{C}^{M\tilde{N}_P\times 1}. \tag{12}$$

$$\hat{\mathbf{W}}'_{\mathrm{MC}}(k) = \left[\hat{\mathbf{W}}_1^T, \hat{\mathbf{W}}_{\mathrm{MC}}^T(k)\right]^T \in \mathbb{C}^{MN_P\times 1} \tag{13}$$

The sub-diagonalized multi-channel Kalman equations for the adaptation of S1 can then be formulated as follows [3, 10]:

$$\mathbf{K}_{\mathrm{MC}}(k) = \frac{R}{M}\mathbf{P}_{\mathrm{MC}}(k)\mathbf{X}_{\mathrm{FxMC}}^H(k)\mathcal{D}_{\mathrm{MC}}^{-1}(k) \in \mathbb{C}^{M\tilde{N}_P\times M} \tag{14}$$

$$\hat{\mathbf{W}}_{\mathrm{MC}}^+(k) = \hat{\mathbf{W}}_{\mathrm{MC}}(k) + \mathbf{K}_{\mathrm{MC}}(k)\mathbf{E}(k) \in \mathbb{C}^{M\tilde{N}_P\times 1} \tag{15}$$

$$\mathbf{P}_{\mathrm{MC}}^+(k) = \mathbf{P}_{\mathrm{MC}}(k) - \frac{R}{M}\mathbf{K}_{\mathrm{MC}}(k)\mathbf{X}_{\mathrm{FxMC}}(k)\mathbf{P}_{\mathrm{MC}}(k) \tag{16}$$

$$\hat{\mathbf{W}}_{\mathrm{MC}}(k+1) = A_{\mathrm{MC}} \cdot \hat{\mathbf{W}}_{\mathrm{MC}}^+(k) \tag{17}$$

$$\mathbf{P}_{\mathrm{MC}}(k+1) = A_{\mathrm{MC}}^2 \cdot \mathbf{P}_{\mathrm{MC}}^+(k) + \boldsymbol{\Psi}_{\Delta\Delta,\mathrm{MC}}(k), \tag{18}$$

$$\mathcal{D}_{\mathrm{MC}}(k) = \frac{R}{M}\mathbf{X}_{\mathrm{FxMC}}(k)\mathbf{P}_{\mathrm{MC}}(k)\mathbf{X}_{\mathrm{FxMC}}^H(k) + \boldsymbol{\Psi}_{ss,\mathrm{MC}}(k). \tag{19}$$

The $M\tilde{N}_P \times M\tilde{N}_P$ covariance matrix $\mathbf{P}_{\mathrm{MC}}(k)$ of the estimation error consists of $\tilde{N}_P^2$ diagonal matrices of size $M \times M$. The forgetting factor A_{MC} of the Kalman algorithm is set to 0.9999. The covariance matrices of the process and measurement noise, $\boldsymbol{\Psi}_{\Delta\Delta,\mathrm{MC}}(k)$ and $\boldsymbol{\Psi}_{ss,\mathrm{MC}}(k)$, are estimated similar to [13]. The reference signals $\mathbf{X}_{\mathrm{FxMC}}(k)$ provided to the Kalman filter of S1 are the nonlinear reference signals $x^p(i)$ filtered by the current estimate $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$ of S2. This prefiltering is necessary because the output of S1 is filtered by $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$ and consequently the resulting error signal $e(i)$ cannot be correlated with the reference of S1 directly. The prefiltering is also known as the *Filtered-x* (Fx) method, e.g., in active noise cancellation [14, 15]. By using the Fx method the adaptation of S1 actually sees a swapped filter order, because it implicitly assumes that the reference is first filtered by $\hat{\mathbf{H}}_{\mathrm{SC}}(k)$ and then by $\hat{\mathbf{W}}_{\mathrm{MC}}(k)$ in the compensation path, too. In the time domain this swapping introduces an error, because the involved filters are time-variant [16]. However, as we are operating in the block-based frequency domain the filter weights are constant during one frame and no error is introduced. Therefore, the $M \times M\tilde{N}_P$ matrix $\mathbf{X}_{\mathrm{FxMC}}(k)$ containing the Fx reference can be calculated with

$$\mathbf{X}_{\mathrm{FxMC}}(k) = [\mathbf{X}_{\mathrm{Fx},3}(k), \mathbf{X}_{\mathrm{Fx},5}(k), \ldots, \mathbf{X}_{\mathrm{Fx},P}(k)] \tag{20}$$

$$\mathbf{X}_{\mathrm{Fx},p}(k) = \mathrm{diag}\left\{\mathbf{G}_{\mathrm{lin}}\mathbf{X}_p(k)\hat{\mathbf{H}}_{\mathrm{SC}}(k)\right\} \in \mathbb{C}^{M\times M}, \tag{21}$$

where $\mathbf{G}_{\mathrm{lin}} = \mathbf{F}_M\mathbf{Q}_{\mathrm{lin}}\mathbf{Q}_{\mathrm{lin}}^H\mathbf{F}_M^{-1}$ and $\mathbf{Q}_{\mathrm{lin}} = (\mathbf{0}_{N_{\mathrm{lin}}-1}\,\mathbf{I}_{M-N_{\mathrm{lin}}+1})^T$ is a zero-padding matrix. It can be shown that the energy scaling factors R/M in (14), (16) and (19) do not need to be changed, neither for the Fx signals nor for the decimated signals in Sec. 3.1.

3.1. Complexity reduction by decimation in frequency

Typically the nonlinear memory of the loudspeaker dynamics is much shorter than the memory of the room impulse response. Consequently the length N_{nl} of the filters $\hat{\boldsymbol{w}}_p(k)$ can be chosen much smaller than the length N_{lin} of $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$. In the frequency domain the spectral weights of $\hat{\boldsymbol{w}}_p(k)$ are smooth and can be represented by fewer FFT bins than $\hat{\boldsymbol{h}}_{\mathrm{rir}}(k)$. This enables a new possibility for reducing the complexity of the multi-channel Kalman filter in S1. The complexity reduction is done by decimating the one-sided spectrum of the Fx reference signals $X'_{\mathrm{FxMC}}(\mu)$ by a factor D according to:

$$\tilde{X}'_{\mathrm{FxMC}}(\tilde{\mu}) = X'_{\mathrm{FxMC}}(\tilde{\mu}D) \text{ for } \tilde{\mu} = 1, 2, \ldots, M/(2D) - 1 \tag{22}$$

where $D = 2^n, n \in \mathbb{N}_0$, μ and $\tilde{\mu}$ are the original and decimated frequency bin indices, respectively. The DC-bin ($\tilde{\mu} = 0$) and Nyquist-bin ($\tilde{\mu} = M/(2D)$) are scaled by $\sqrt{D}$ to prevent them from dominating the decimated signal. After decimation the full spectrum is reconstructed exploiting the complex-conjugate symmetry. The spectrum of the error signal $E(\mu)$ is treated in the same way. Now the Kalman equations (14)–(19) can be computed at the reduced FFT size of $\tilde{M} = M/D$. The decimated FFT size $\tilde{M}$ must satisfy $\tilde{M} \geq N_{\mathrm{nl}} + 1$ to be able to represent all coefficients of $\hat{\boldsymbol{w}}_p(i)$. In order to stay in the same overlap-save framework, the computed weights of the decimated S1 are transformed into the time domain, zero-padded to a length of M and transformed back to the frequency domain with an FFT of size M. To further reduce the complexity the constraining in (21) and (9) can be omitted by setting $\mathbf{G}_{\mathrm{lin,nl}} = \frac{M-N_{\mathrm{lin,nl}}+1}{M}\mathbf{I}_M$, [12]. The complexity of the linear Kalman $\mathcal{C}_{\mathrm{lin}}$, the Full-MCK $\mathcal{C}_{\mathrm{F}}$, the decimated DualStage-MCK $\mathcal{C}_{\mathrm{DS}}^*$ without constraining, and the decimated DualStage-MCK $\mathcal{C}_{\mathrm{DS}}$ with constraining is calculated by counting the number of real additions and multiplications for each frame as in [3]

$$\mathcal{C}_{\mathrm{lin}} = \mathcal{O}(M) + \mathcal{O}(M\log_2 M)$$

$$\mathcal{C}_{\mathrm{F}} = \mathcal{O}(N_P^3 M) + \mathcal{O}(N_P^2 M) + \mathcal{O}(N_P M) + \mathcal{O}(N_P M\log_2 M)$$

$$\mathcal{C}_{\mathrm{DS}}^* = \mathcal{O}(\tilde{N}_P^3\tilde{M}) + \mathcal{O}(\tilde{N}_P^2\tilde{M}) + \mathcal{O}(\tilde{N}_P\tilde{M}) + \mathcal{O}(\tilde{N}_P\tilde{M}\log_2\tilde{M}) + \mathcal{C}_{\mathrm{lin}} + \mathcal{C}_{\mathrm{Fx}}, \tag{23}$$

$$\mathcal{C}_{\mathrm{DS}} = \mathcal{C}_{\mathrm{DS}}^* + \mathcal{C}_{\mathrm{constr}}, \tag{24}$$

where $\mathcal{C}_{\mathrm{Fx}} = \mathcal{O}(\tilde{N}_P M)$ and $\mathcal{C}_{\mathrm{constr}} = \mathcal{O}((\tilde{N}_P + 1)M\log_2 M)$ is the overhead for the application of the Fx-weights and the additional constraining, respectively. These approximations will be used in Sec. 4.2 to evaluate the complexity reduction obtained by the DualStage-MCK. It has to be noted that the far-end signal $\mathbf{X}_{\mathrm{MC}}(k)$ is not necessarily smooth, even if it is constrained after the Fx stage. Therefore, the decimation of $\mathbf{X}_{\mathrm{FxMC}}(k)$ introduces an error. However, as the sub-diagonalized Kalman equations adapt the spectral weights in the reduced FFT domain separately for each FFT bin, the adaptation of the bins of interest is still correct. Note that the decimated version of $\mathbf{X}_{\mathrm{FxMC}}(k)$ is never converted back to the time domain. This would, of course, introduce quite severe alias artifacts. The same holds for the decimated version of the error signal $\mathbf{E}(k)$.

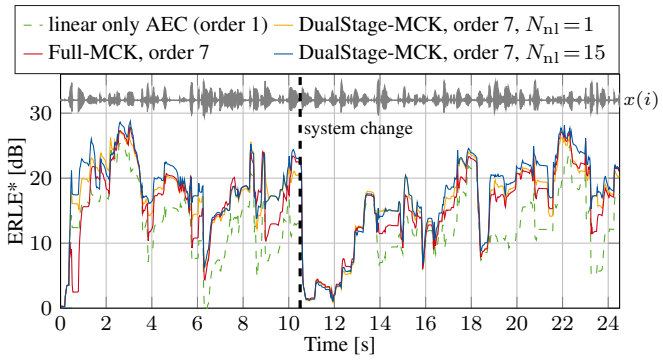
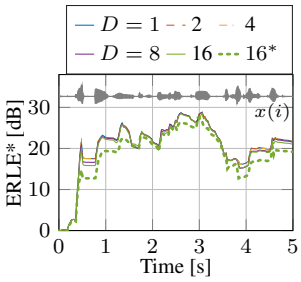


Fig. 3. DualStage-MCK ($D=1$) vs. Full-MCK, abrupt change of acoustic path at $t=10.5\,\mathrm{s}$ from 'on microphone stand' to 'on desk'.



(a)

D	$\frac{\mathcal{C}_{\mathrm{DS}}}{\mathcal{C}_{\mathrm{F}}}$	$\frac{\mathcal{C}^*_{\mathrm{DS}}}{\mathcal{C}_{\mathrm{F}}}$	$\frac{\mathcal{C}_{\mathrm{DS}}}{\mathcal{C}_{\mathrm{lin}}}$	$\frac{\mathcal{C}^*_{\mathrm{DS}}}{\mathcal{C}_{\mathrm{lin}}}$
1	0.92	0.65	11.89	8.33
2	0.64	0.36	8.22	4.67
4	0.50	0.23	6.47	2.92
8	0.44	0.16	5.64	2.08
16	0.41	0.13	5.24	1.69

(b)

Fig. 4. ERLE* for the DualStage-MCK ($N_{\mathrm{nl}}=15$) (a) and relative complexity (b) for different decimation factors D. ---16* and $\mathcal{C}^*_{\mathrm{DS}}$ indicate the performance and complexity without constraining, respectively.

4. EVALUATION

We evaluate the DualStage-MCK on measured signals, for which the true nonlinear system is not known. For the acquisition of nonlinear distorted echo signals, that are close to a real world scenario, a Class D amplifier, an actual smartphone loudspeaker and a digital MEMS microphone was used. All components were integrated into a smartphone mockup with dimensions of $14\,\mathrm{cm} \times 7\,\mathrm{cm} \times 1\,\mathrm{cm}$. The echo-to-noise-ratio (ENR) of the recordings was $39\,\mathrm{dB}$. The mockup was placed onto a microphone stand inside a studio booth to have a controlled acoustic environment with low reverberation ($T_{60}=0.12\,\mathrm{s}$). For the evaluation of abrupt changes in the acoustic path in Fig. 3, the mockup was also placed on a desk inside the studio booth. For the far-end signal speech samples from the TSP corpus [17] with different genders and speakers are used to account for varying correlation properties of the excitation signal. As state-of-the-art anchors we present results for the Full-MCK [3] and the cascaded approach without memory [7], where the latter is mimicked with the DualStage-MCK using $N_{\mathrm{nl}}=1$ samples. If nothing else is specified we use a sampling rate of $8\,\mathrm{kHz}$, a frame and FFT size of $M=256\,\mathrm{samples}$, a frame shift of $R=64\,\mathrm{samples}$ and an odd order power series of order 7 for all algorithms. For the proposed DualStage-MCK with memory we use $N_{\mathrm{nl}}=15$ samples. These values were chosen, because our experiments showed that a further increase of both, nonlinear order and memory, does not significantly improve the performance. When dealing with measured signals only $x(i)$ and $y(i)$ are accessible. Hence, we must rely on the estimated echo return loss enhancement (ERLE^*) for evaluation, that is defined by $\mathrm{ERLE}^*(i)[\mathrm{dB}] = 10\log_{10}\left(\mathrm{E}\{y^2(i)\}/\mathrm{E}\{e^2(i)\}\right)$, where the expectation operator $\mathrm{E}\{\cdot\}$ is realized by recursive averaging.

4.1. Convergence behavior

In order to simulate a time-variance in the system, we introduced an abrupt change in the acoustics at $t=10.5\,\mathrm{s}$ by switching between the recordings obtained for the mockup on the microphone stand and on the desk. Fig. 3 shows the resulting ERLE^* for the proposed DualStage-MCK with memory ($N_{\mathrm{nl}}=15$), the DualStage-MCK without memory ($N_{\mathrm{nl}}=1$) and for the Full-MCK. For all times the DualStage-MCK with memory of order 7 consistently outperforms the linear only AEC (DualStage-MCK, order 1) by several dB. From $0\,\mathrm{s}$ to $2\,\mathrm{s}$ it can clearly be seen that the DualStage-MCK with memory shows fastest convergence, followed by the DualStage-MCK without memory and the Full-MCK. Furthermore, both DualStage-MCK variants do not lose speed compared to the linear only AEC. Right after the path change at $t=10.5\,\mathrm{s}$ all algorithms show a similar convergence behavior. From $t=14\,\mathrm{s}$ the DualStage-MCK algorithms outperform the Full-MCK and for $t>16\,\mathrm{s}$ the proposed DualStage-MCK with memory outperforms the DualStage-MCK without memory. For all algorithms the convergence speed at the beginning of the recording is faster than after the abrupt change at $t=10.5\,\mathrm{s}$. This is because the first order Markov model, assumed by the Kalman filter for the time variance of the echo path, cannot model abrupt changes. Therefore, all Kalman equations assume a more or less converged state and only use a relatively small stepsize after $t=10.5\,\mathrm{s}$, too.

4.2. Complexity reduction

Up to now the computational complexity is quite high as a multi-channel Kalman filter is used in S1 and, in addition, a single-channel Kalman filter in S2. Both algorithms are running with an FFT-size of $M=256$. As already shown in Sec. 3.1, the FFT-size for S1 can be chosen much smaller than for S2, because the $\hat{\boldsymbol{w}}_p$ are short. We are using $N_{\mathrm{nl}}=15$ and hence an FFT-size of $\tilde{M}=16$ should be sufficient for the multi-channel Kalman filter in S1 to adapt $\hat{\boldsymbol{w}}_p$. This corresponds to a maximum decimation factor of $D=M/\tilde{M}=16$. Fig. 4 (a) shows the performance of the DualStage-MCK for different decimation factors D. No significant decrease of the ERLE^* performance can be seen when increasing the decimation factor. As shown by the dashed green line (---) the performance slightly decreases if the constraining in (21) and (9) is omitted. We can conclude from Fig. 4 (b) that with decimation and constraining a reduction of the complexity by a factor of 2.3 is possible without degrading the echo reduction performance. If the constraining is omitted the total reduction of performance amounts to a factor of 7 with $D=16$. Then, the complexity of the DualStage-MCK with memory is only 69 % higher than the complexity of the linear only AEC. This underlines the relevance of the DualStage-MCK with memory for the application on mobile platforms with limited resources.

5. CONCLUSION

We presented the dual-stage multi-channel Kalman filter for nonlinear AEC that mimics the cascaded nature of the true echo path by serially concatenating one stage modelling the nonlinear loudspeaker with memory and a subsequent stage modelling the acoustic transmission through the room. By using the filtered-x method and the multi-channel Kalman algorithm in the block frequency domain for adaptation, a near-optimum stepsize control for the cascaded structure is realized. The results for dynamic acoustic scenarios confirm the superior performance of the proposed method compared to state-of-the-art solutions. Furthermore by exploiting the short length of the nonlinear memory, the computational complexity of the proposed solution could significantly be reduced by a decimation of the FFT size. This makes the proposed solution attractive for real-time speech communication in mobile devices.

6. REFERENCES

[1] P. Shah, I. Lewis, S. Grant, and S. Angrignon, "Nonlinear acoustic echo cancellation using feedback," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 610–614.

[2] Fabian Kuech and Walter Kellermann, "Orthogonalized power filters for nonlinear acoustic echo cancellation," *Signal Processing*, vol. 86, no. 6, pp. 1168–1181, June 2006.

[3] Sarmad Malik and Gerald Enzner, "State-Space Frequency-Domain Adaptive Filtering for Nonlinear Acoustic Echo Cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2065–2079, Sept. 2012.

[4] Christian Hofmann, Christian Huemmer, Michael Guenther, and Walter Kellermann, "Significance-aware filtering for nonlinear acoustic echo cancellation," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 113, Nov. 2016.

[5] Bryan S Nollett and Douglas L Jones, "Nonlinear echo cancellation for hands-free speakerphones," *IEEE Workshop on Nonlinear Signal and Image Processing*, p. 5, 1997.

[6] Alexander Stenger and Walter Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, no. 9, pp. 1747–1760, Sept. 2000.

[7] S. Malik and G. Enzner, "A Variational Bayesian Learning Approach for Nonlinear Acoustic Echo Control," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5853–5867, Dec. 2013.

[8] Moctar I. Mossi, Christelle Yemdji, Nicholas Evans, Christophe Beaugeant, and Philippe Degry, "Robust and low-cost cascaded non-linear acoustic echo cancellation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 89–92, IEEE.

[9] Muhammad Z. Ikram, "Non-linear acoustic echo cancellation using cascaded Kalman filtering," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 1320–1324, IEEE.

[10] S. Kühl, C. Antweiler, T. Hübschen, and P. Jax, "Kalman filter based stereo system identification with auto- and cross-decorrelation," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, Mar. 2017, pp. 181–185.

[11] A. Guerin, G. Faucon, and R. Le Bouquin-Jeannes, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 672–683, Nov. 2003.

[12] Jacob Benesty, Tomas Gänsler, Dennis R. Morgan, M. Mohan Sondhi, and Steven L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Digital Signal Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[13] Gerald Enzner, *A Model-Based Optimum Filtering Approach to Acoustic Echo Control: Theory and Practice*, Dissertation, IND, RWTH Aachen, Aachen, Germany, Apr. 2006.

[14] S. M. Kuo and D. R. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–973, June 1999.

[15] Constantin Paleologu, Jacob Benesty, and Silviu Ciochină, "Adaptive filtering for the identification of bilinear forms," *Digital Signal Processing*, vol. 75, pp. 153–167, Apr. 2018.

[16] S. Kühl, C. Antweiler, T. Hübschen, and P. Jax, "Kalman filter based system identification exploiting the decorrelation effects of linear prediction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 4790–4794.

[17] Peter Kabal, "TSP speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.