A Synergistic Kalman- and Deep Postfiltering Approach to Acoustic Echo Cancellation

Thomas Haubner, Mhd. Modar Halimeh, Andreas Brendel, and Walter Kellermann
Multimedia Communications and Signal Processing, Friedrich-Alexander-University Erlangen-Nürnberg,
Cauerstr. 7, D-91058 Erlangen, Germany, e-mail: thomas.haubner@fau.de

***Abstract*—We introduce a synergistic approach to double-talk robust acoustic echo cancellation combining adaptive Kalman filtering with a deep neural network-based postfilter. The proposed algorithm overcomes the well-known limitations of Kalman filter-based adaptation control in scenarios characterized by abrupt echo path changes. As the key innovation, we suggest to exploit the different statistical properties of the interfering signal components for robustly estimating the adaptation step size. This is achieved by leveraging the postfilter near-end estimate and the estimation error of the Kalman filter. The proposed scheme allows for rapid reconvergence of the adaptive filter after abrupt echo path changes without compromising the steady-state performance achieved by state-of-the-art approaches in static scenarios.**

***Index Terms*—Adaptation Control, Acoustic Echo Cancellation, Kalman Filter, Post Filter, Echo Path Change**

I. Introduction

Acoustic echo cancellation (AEC) is an essential part of any full-duplex hands-free acoustic communication application, e.g., teleconferencing or human-machine dialogue systems [1]. Research on AEC algorithms has evolved from simple time-domain adaptive filter (AF) approaches [2] and computationally efficient frequency-domain implementations [3] to recent deep learning-based approaches [4], [5].

In general, current AEC algorithms can be classified into model-based AF-postfilter (PF) approaches and direct deep learning-based PF (DPF) approaches. While model-based algorithms show unmatched generalization to unknown acoustic environments, they require sophisticated adaptation control mechanisms to cope with double-talk situations [1]. In particular the probabilistically motivated inference of the AF coefficients by a Kalman Filter (KF) [6], [7] enabled the much sought-after continuous adaptation control without the need of a double-talk detector. However, despite the increased double-talk robustness, KF approaches suffer from slow reconvergence after abrupt Echo Path Changes (EPCs) which are commonly encountered with portable devices [8], [9]. This slow recovery is caused by overestimating the noise Power Spectral Density (PSD) matrix and often remedied by auxiliary mechanisms like shadow filters [8] or trained noise models [10] which require additional computational power. Unlike AFs, DPF algorithms for AEC [4], [5] do not require online adaptation when trained on adequate datasets. However, as communication devices are usually exposed to a large variety of acoustic environments, Deep Neural Network (DNN) models with many parameters need to be trained which limits their applicability to devices with sufficient computational power. These computational requirements can be mitigated by using smaller networks with less parameters when combining model-based and deep learning approaches [11], [12], [13]. However, most methods treat the AF estimation independently from the DPF. Recently, a DNN-supported Expectation-Maximization (EM) optimization of a local Gaussian model has shown improved performance for joint reverberation, echo and noise reduction [14]. However, a narrowband assumption is made and the filter estimates are assumed to be time-invariant which limits the performance for time-varying scenarios [14].

In this paper, we introduce a synergistic approach which combines a broadband adaptive KF and a DPF and thereby successfully copes with time-varying acoustic environments. We show how the slow reconvergence of the KF after abrupt EPCs can be remedied by exploiting the different signal statistics of the various interfering components. By efficiently fusing the DPF near-end estimate and the KF estimation error, a robust estimate for the KF step size is obtained without any auxiliary mechanisms.

In the following, we use bold uppercase letters for matrices and bold lowercase letters for vectors with underlined symbols indicating time-domain quantities. A matrix element in the mth row and the nth column is indicated by $[\cdot]_{mn}$. We denote the all-zero matrix of dimensions $D_1 \times D_2$ by $\mathbf{0}_{D_1 \times D_2}$, the D-dimensional identity matrix by $\boldsymbol{I}_D$ and the D-dimensional Discrete Fourier Transform (DFT) matrix by $\boldsymbol{F}_D$. Furthermore, the transposition and Hermitian transposition are represented by $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$, respectively. The proper complex Gaussian Probability Density Function (PDF), with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Psi}$, is denoted by $\mathcal{N}_c(\cdot|\boldsymbol{\mu}, \boldsymbol{\Psi})$ and the expectation operator by $\mathbb{E}[\cdot]$. Finally, the diag$(\cdot)$ operator constructs a diagonal matrix from its vector-valued argument.

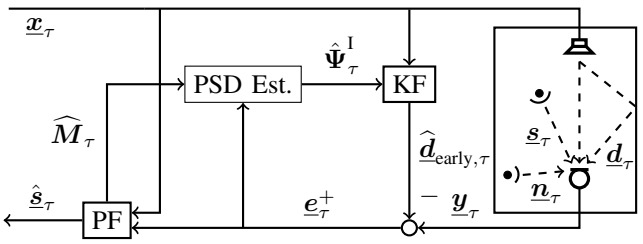

Fig. 1. Proposed synergistic KF+DPF approach to AEC.

This work was supported by the German Research Foundation - 282835863 - within the Research Unit FOR2457 "Acoustic Sensor Networks.

II. Probabilistic Signal Model

The observed time-domain microphone signal block $\underline{\boldsymbol{y}}_\tau$ is modelled as a linear superposition of an early echo component $\underline{\boldsymbol{d}}_{\text{early},\tau}$, a late echo component $\underline{\boldsymbol{d}}_{\text{late},\tau}$, background noise $\underline{\boldsymbol{n}}_\tau$ and a near-end speaker $\underline{\boldsymbol{s}}_\tau$ (cf. Fig. 1), as follows:

$$\underline{\boldsymbol{y}}_\tau = \underline{\boldsymbol{d}}_{\text{early},\tau} + \underline{\boldsymbol{d}}_{\text{late},\tau} + \underline{\boldsymbol{n}}_\tau + \underline{\boldsymbol{s}}_\tau \in \mathbb{R}^R \tag{1}$$

where signal block $\underline{\boldsymbol{y}}_\tau$ at time index τ consists of R consecutive samples, i.e.,

$$\underline{\boldsymbol{y}}_\tau = \left[\underline{y}_{\tau R-R+1}, \underline{y}_{\tau R-R+2}, \ldots, \underline{y}_{\tau R}\right]^{\mathrm{T}} \in \mathbb{R}^R, \tag{2}$$

and $\underline{\boldsymbol{d}}_{\text{early},\tau}$, $\underline{\boldsymbol{d}}_{\text{late},\tau}$, $\underline{\boldsymbol{s}}_\tau$ and $\underline{\boldsymbol{n}}_\tau$ are defined analogously. The early echo component $\underline{\boldsymbol{d}}_{\text{early},\tau}$ is modelled by a linear convolution of a Finite Impulse Response (FIR) filter $\underline{\boldsymbol{w}}_\tau \in \mathbb{R}^L$ with the corresponding far-end signal block $\underline{\boldsymbol{x}}_\tau \in \mathbb{R}^{L+R-1}$. The linear convolution can efficiently be implemented by a Partitioned-Block Convolution (PBC). For this the FIR filter $\underline{\boldsymbol{w}}_\tau$ is separated into $B = \frac{L}{R}$ partitions $\underline{\boldsymbol{w}}_{b,\tau} \in \mathbb{R}^R$. Subsequently, each partition is convolved with a corresponding delayed far-end block

$$\underline{\boldsymbol{x}}_{b,\tau} = \left[\underline{x}_{(\tau-b)R-M+1}, \underline{x}_{(\tau-b)R-M+2}, \ldots, \underline{x}_{(\tau-b)R}\right]^{\mathrm{T}} \in \mathbb{R}^M \tag{3}$$

of length $M = 2R$ and the convolution products are added. By implementing each linear convolution in the DFT domain one obtains [7]:

$$\underline{\boldsymbol{d}}_{\text{early},\tau} = \sum_{b=0}^{B-1} \boldsymbol{Q}_1^{\mathrm{T}} \boldsymbol{F}_M^{-1} \boldsymbol{X}_{b,\tau} \boldsymbol{w}_{b,\tau} \in \mathbb{R}^R \tag{4}$$

with the constraint matrix $\boldsymbol{Q}_1^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{0}_{R\times R} & \boldsymbol{I}_R\end{bmatrix}$, the DFT-domain FIR filter partition $\boldsymbol{w}_{b,\tau} = \boldsymbol{F}_M \boldsymbol{Q}_2 \underline{\boldsymbol{w}}_{b,\tau} \in \mathbb{C}^M$, the DFT-domain far-end signal block $\boldsymbol{X}_{b,\tau} = \operatorname{diag}\left(\boldsymbol{F}_M \underline{\boldsymbol{x}}_{b,\tau}\right)$ and the zero-padding matrix $\boldsymbol{Q}_2^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{I}_R & \boldsymbol{0}_{R\times R}\end{bmatrix}$. A relation between the noisy observation $\underline{\boldsymbol{y}}_\tau$ and the filter partitions $\boldsymbol{w}_{b,\tau}$ is obtained by inserting the PBC model (4) into the signal model (1). By using $\boldsymbol{y}_\tau = \boldsymbol{F}_M \boldsymbol{Q}_1 \underline{\boldsymbol{y}}_\tau$ this time-domain observation equation is transformed to the DFT domain:

$$\boldsymbol{y}_\tau = \sum_{b=0}^{B-1} \boldsymbol{C}_{b,\tau} \boldsymbol{w}_{b,\tau} + \boldsymbol{d}_{\text{late},\tau} + \boldsymbol{n}_\tau + \boldsymbol{s}_\tau \in \mathbb{C}^M \tag{5}$$

with the DFT-domain signal components $\boldsymbol{d}_{\text{late},\tau}$, $\boldsymbol{n}_\tau$, and $\boldsymbol{s}_\tau$ and the overlap-save-constrained far-end signal blocks $\boldsymbol{C}_{b,\tau} = \boldsymbol{F}_M \boldsymbol{Q}_1 \boldsymbol{Q}_1^{\mathrm{T}} \boldsymbol{F}_M^{-1} \boldsymbol{X}_{b,\tau}$. Note that the corresponding time-domain signals can be computed by the inverse transform $\underline{\boldsymbol{y}}_\tau = \boldsymbol{Q}_1^{\mathrm{T}} \boldsymbol{F}_M^{-1} \boldsymbol{y}_\tau$. The late echo component $\boldsymbol{d}_{\text{late},\tau}$, the background noise $\boldsymbol{n}_\tau$, and the near-end speaker $\boldsymbol{s}_\tau$ are considered as additive disturbances when estimating the AF partitions $\boldsymbol{w}_{b,\tau}$ in Eq. (5). In the following, we model each of these interfering components as a zero-mean, non-stationary and spectrally uncorrelated proper complex Gaussian random process by the respective PDFs

$$p(\boldsymbol{d}_{\text{late},\tau}) = \mathcal{N}_c(\boldsymbol{d}_{\text{late},\tau} | \boldsymbol{0}_{M\times 1}, \boldsymbol{\Psi}_\tau^{\mathrm{D}_{\text{late}}}) \tag{6}$$

$$p(\boldsymbol{n}_\tau) = \mathcal{N}_c(\boldsymbol{n}_\tau | \boldsymbol{0}_{M\times 1}, \boldsymbol{\Psi}_\tau^{\mathrm{N}}), \tag{7}$$

$$p(\boldsymbol{s}_\tau) = \mathcal{N}_c(\boldsymbol{s}_\tau | \boldsymbol{0}_{M\times 1}, \boldsymbol{\Psi}_\tau^{\mathrm{S}}), \tag{8}$$

with the diagonal PSD matrices $\boldsymbol{\Psi}_\tau^{\mathrm{D}_{\text{late}}}$, $\boldsymbol{\Psi}_\tau^{\mathrm{N}}$ and $\boldsymbol{\Psi}_\tau^{\mathrm{S}} \in \mathbb{C}^{M\times M}$. We assume $\boldsymbol{d}_{\text{late},\tau}$, $\boldsymbol{n}_\tau$ and $\boldsymbol{s}_\tau$ to be mutually uncorrelated

$$\mathbb{E}\left[\boldsymbol{d}_{\text{late},\tau} \boldsymbol{n}_\tau^{\mathrm{H}}\right] = \mathbb{E}\left[\boldsymbol{d}_{\text{late},\tau} \boldsymbol{s}_\tau^{\mathrm{H}}\right] = \mathbb{E}\left[\boldsymbol{s}_\tau \boldsymbol{n}_\tau^{\mathrm{H}}\right] = \boldsymbol{0}_{M\times M}. \tag{9}$$

Finally, to account for the time-variance of acoustic environments, the temporal evolution of each AF partition $\boldsymbol{w}_{b,\tau}$ is modelled by a DFT-domain random walk Markov model [7]

$$\boldsymbol{w}_{b,\tau} = A\ \boldsymbol{w}_{b,\tau-1} + \Delta\boldsymbol{w}_{b,\tau} \tag{10}$$

with the process noise vector of the bth partition $\Delta\boldsymbol{w}_{b,\tau}$ and the state transition coefficient $0 < A < 1$. The process noise $\Delta\boldsymbol{w}_{b,\tau}$ is assumed to be distributed according to

$$p(\Delta\boldsymbol{w}_{b,\tau}) = \mathcal{N}_c(\Delta\boldsymbol{w}_{b,\tau} | \boldsymbol{0}_{M\times 1}, \boldsymbol{\Psi}_{b,\tau}^{\Delta\mathrm{W}}) \tag{11}$$

with the diagonal process noise PSD matrix $\boldsymbol{\Psi}_{b,\tau}^{\Delta\mathrm{W}}$. Note that Eqs. (5) - (11) represent a linear Gaussian state-space model with the DFT-domain AF partitions $\boldsymbol{w}_{b,\tau}$ as the states and the microphone signal blocks $\boldsymbol{y}_\tau$ as the observations.

III. Acoustic Echo Cancellation Model

The considered AEC architecture is depicted in Fig. 1. A linear AF estimates the early echo component $\underline{\boldsymbol{d}}_{\text{early},\tau}$ which is then subtracted from the noisy observation $\underline{\boldsymbol{y}}_\tau$. The estimation error signal $\underline{\boldsymbol{e}}_\tau^+$ is used as input to a DPF. Finally, for a double-talk robust adaptation control of the AF partitions $\boldsymbol{w}_{b,\tau}$, we propose a novel approach to estimate the observation noise PSD matrix by exploiting the near-end speaker estimate of the DPF and the estimation error $\underline{\boldsymbol{e}}_\tau^+$.

A. Adaptive Kalman Filter

We model the state posterior of each AF partition $\boldsymbol{w}_{b,\tau}$, given all preceding observations $\boldsymbol{Y}_{1:\tau} = \begin{bmatrix}\boldsymbol{y}_1, & \ldots, & \boldsymbol{y}_\tau\end{bmatrix}$, by

$$p(\boldsymbol{w}_{b,\tau} | \boldsymbol{Y}_{1:\tau}) = \mathcal{N}_c\left(\boldsymbol{w}_{b,\tau} | \hat{\boldsymbol{w}}_{b,\tau}, \boldsymbol{P}_{b,\tau}\right) \tag{12}$$

with mean $\hat{\boldsymbol{w}}_{b,\tau}$ and diagonal state uncertainty matrix $\boldsymbol{P}_{b,\tau}$. Due to the linear Gaussian model (cf. Eqs. (5) - (11)), a closed-form inference of the state posterior is given by the KF equations. By setting the transition factor to one in the prediction of the echo and introducing a gradient constraint, the diagonalized Partitioned-Block KF (PBKF) is obtained [7]

$$\begin{aligned}
\widehat{\boldsymbol{d}}_{\text{early},\tau} &= \sum_{b=0}^{B-1} \boldsymbol{C}_{b,\tau} \hat{\boldsymbol{w}}_{b,\tau-1} \approx A \sum_{b=0}^{B-1} \boldsymbol{C}_{b,\tau} \hat{\boldsymbol{w}}_{b,\tau-1} \\
\boldsymbol{e}_\tau^+ &= \boldsymbol{y}_\tau - \widehat{\boldsymbol{d}}_{\text{early},\tau} \\
\boldsymbol{P}_{b,\tau-1}^+ &= A^2\ \boldsymbol{P}_{b,\tau-1} + \boldsymbol{\Psi}_{b,\tau}^{\Delta\mathrm{W}} \\
\boldsymbol{\Lambda}_{b,\tau} &= \boldsymbol{P}_{b,\tau-1}^+ \left(\sum_{\tilde{b}=0}^{B-1} \boldsymbol{X}_{\tilde{b},\tau} \boldsymbol{P}_{\tilde{b},\tau-1}^+ \boldsymbol{X}_{\tilde{b},\tau}^{\mathrm{H}} + \frac{M}{R} \boldsymbol{\Psi}_\tau^{\mathrm{I}}\right)^{-1} \\
\hat{\boldsymbol{w}}_{b,\tau} &= \hat{\boldsymbol{w}}_{b,\tau-1} + \boldsymbol{G} \boldsymbol{\Lambda}_{b,\tau} \boldsymbol{X}_{b,\tau}^{\mathrm{H}} \boldsymbol{e}_\tau^+ \\
\boldsymbol{P}_{b,\tau} &= \left(\boldsymbol{I}_M - \frac{R}{M} \boldsymbol{\Lambda}_{b,\tau} \boldsymbol{X}_{b,\tau}^{\mathrm{H}} \boldsymbol{X}_{b,\tau}\right) \boldsymbol{P}_{b,\tau-1}^+
\end{aligned} \tag{13}$$

with the estimated early echo signal $\widehat{\boldsymbol{d}}_{\text{early},\tau}$, the prior error $\boldsymbol{e}_\tau^+$, the gradient constraint matrix $\boldsymbol{G} = \boldsymbol{F}_M \boldsymbol{Q}_2 \boldsymbol{Q}_2^{\mathrm{T}} \boldsymbol{F}_M^{-1}$ and

the adaptive diagonal step size matrix $\boldsymbol{\Lambda}_{b,\tau}$. The double-talk robustness and convergence properties of the KF crucially depend on a precise estimation of the observation noise PSD matrix $\boldsymbol{\Psi}_\tau^{\mathrm{I}} = \boldsymbol{\Psi}_\tau^{\mathrm{D_{late}}} + \boldsymbol{\Psi}_\tau^{\mathrm{N}} + \boldsymbol{\Psi}_\tau^{S}$ (cf. Sec. IV).

B. Deep Neural Network-based Postfilter

The aim of any PF is the estimation of the near-end signal $\underline{\boldsymbol{s}}_\tau$ (cf. Eq. (1)) given the estimated error $\underline{\boldsymbol{e}}_\tau^+ = \boldsymbol{Q}_1^{\mathrm{T}} \boldsymbol{F}_M^{-1} \boldsymbol{e}_\tau^+$ (cf. Eq. (13)). We consider a recurrent DNN-based PF which is inspired by [12], [15] and comprises four layers. The first layer is a dense feedforward layer with tanh activations which combines the input features to a vector of dimension P. Subsequently, two stacked gated recurrent unit (GRU) layers are added which extract temporal information from the combined features. Finally, a dense feedforward output layer with sigmoid activations is used to map the GRU states to a corresponding frame-wise diagonal masking matrix $\widehat{\boldsymbol{M}}_\tau$.

As input features $\tilde{\boldsymbol{u}}_{\mathrm{feat},\tau}$ to the DNN, we use the normalized logarithmic power spectra of the prior error and the far-end signal. For this, we first compute the Short-Time Fourier Transform (STFT) of the time-domain signals (cf. Sec. II),

$$\tilde{\boldsymbol{u}}_\tau = \begin{bmatrix} \tilde{\boldsymbol{e}}_\tau^+ \\ \tilde{\boldsymbol{x}}_{0,\tau} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_M \boldsymbol{V} \left[(\underline{\boldsymbol{e}}_{\tau-1}^+)^{\mathrm{T}} \quad (\underline{\boldsymbol{e}}_\tau^+)^{\mathrm{T}} \right]^{\mathrm{T}} \\ \boldsymbol{F}_M \boldsymbol{V} \underline{\boldsymbol{x}}_{0,\tau} \end{bmatrix} \in \mathbb{C}^{2M} \quad (14)$$

with the diagonal window matrix $\boldsymbol{V} \in \mathbb{R}^{M \times M}$ and $\tilde{(\cdot)}$ denoting windowed STFT-domain quantities. Subsequently, the normalized logarithmic power spectrum is computed by

$$[\tilde{\boldsymbol{u}}_{\mathrm{feat},\tau}]_m = \frac{\log(\max(|[\tilde{\boldsymbol{u}}_\tau]_m|^2, \epsilon_1)) - [\boldsymbol{\mu}]_m}{[\boldsymbol{\sigma}]_m}, \quad (15)$$

with $m = 0, \ldots, 2M-1$ and $\epsilon_1 > 0$ being a small number to avoid numerical instabilities. Here, the estimated mean and standard deviation of the feature vector are denoted by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, respectively. Note that due to the symmetry of the STFT only the non-redundant $M+2$ frequency bins of $\tilde{\boldsymbol{u}}_{\mathrm{feat},\tau}$ are used as features. The parameters $\boldsymbol{\theta}$ of the DNN are trained by minimizing the cost function [14], [16]

$$\mathcal{J}_{\mathrm{PF}}(\boldsymbol{\theta}) = \sum_{\tau,m} d_{\mathrm{KL}}(|\tilde{s}_{m\tau}|, |\hat{\tilde{s}}_{m\tau}|), \quad (16)$$

defined by the Kullback-Leibler divergence

$$d_{\mathrm{KL}}(|\tilde{s}_{m\tau}|, |\hat{\tilde{s}}_{m\tau}|) = -|\tilde{s}_{m\tau}| \log(|\hat{\tilde{s}}_{m\tau}| + \epsilon_2) + |\hat{\tilde{s}}_{m\tau}| \quad (17)$$

between the magnitudes of the true STFT near-end component $\tilde{s}_{m\tau} = \left[\boldsymbol{F}_M \boldsymbol{V} \left[(\underline{\boldsymbol{s}}_{\tau-1})^{\mathrm{T}} \quad (\underline{\boldsymbol{s}}_\tau)^{\mathrm{T}} \right]^{\mathrm{T}} \right]_m$ and the estimated one $\hat{\tilde{s}}_{m\tau} = [\widehat{\boldsymbol{M}}_\tau]_{mm} [\tilde{\boldsymbol{e}}_\tau^+]_m$. All constant terms have been omitted in Eq. (17) and a regularization term ϵ_2 has been included [16]. Note that any mask-based PF could be used to support the observation noise PSD estimation of the KF (cf. Sec. IV).

IV. Proposed Power Spectral Density Estimation

For a fast-converging and double-talk robust adaptation of the AF partitions $\hat{\boldsymbol{w}}_{b,\tau}$, a precise estimation of the observation noise PSD matrix $\boldsymbol{\Psi}_\tau^{\mathrm{I}}$ and process noise PSD matrices $\boldsymbol{\Psi}_{b,\tau}^{\Delta \mathrm{W}}$ is decisive. In particular, the observation noise PSD matrix $\boldsymbol{\Psi}_\tau^{\mathrm{I}}$ is, due to its fast changing behaviour, difficult to estimate. In contrast to state-of-the-art approaches, we suggest to exploit the different statistical properties of the signal components generating the observation $\boldsymbol{y}_\tau$ (cf. Eqs. (1) and (5)).

We start by representing the prior error signal

$$\boldsymbol{e}_\tau^+ = \boldsymbol{e}_{\mathrm{early},\tau}^+ + \boldsymbol{p}_\tau + \boldsymbol{s}_\tau \quad (18)$$

in terms of the early echo error signal $\boldsymbol{e}_{\mathrm{early},\tau}^+ = \boldsymbol{d}_{\mathrm{early},\tau} - \widehat{\boldsymbol{d}}_{\mathrm{early},\tau}$, the late reverberation and background noise signal $\boldsymbol{p}_\tau = \boldsymbol{d}_{\mathrm{late},\tau} + \boldsymbol{n}_\tau$ and the desired near-end speaker signal $\boldsymbol{s}_\tau$. By assuming the early echo error $\boldsymbol{e}_{\mathrm{early},\tau}^+$, the noise signal $\boldsymbol{p}_\tau$ and the near-end speaker signal $\boldsymbol{s}_\tau$ to be mutually uncorrelated, the PSD matrix of the prior error $\boldsymbol{e}_\tau^+$ is given by

$$\boldsymbol{\Psi}_\tau^{\mathrm{E}} = \boldsymbol{\Psi}_\tau^{\mathrm{E_{early}}} + \boldsymbol{\Psi}_\tau^{\mathrm{P}} + \boldsymbol{\Psi}_\tau^{\mathrm{S}} = \boldsymbol{\Psi}_\tau^{\mathrm{E_{early}}} + \boldsymbol{\Psi}_\tau^{\mathrm{I}} \quad (19)$$

with $\boldsymbol{\Psi}_\tau^{\mathrm{P}} = \boldsymbol{\Psi}_\tau^{\mathrm{D_{late}}} + \boldsymbol{\Psi}_\tau^{\mathrm{N}}$ and $\boldsymbol{\Psi}_\tau^{\mathrm{E_{early}}} = \mathbb{E}\left[\boldsymbol{e}_{\mathrm{early},\tau}^+ \left(\boldsymbol{e}_{\mathrm{early},\tau}^+ \right)^{\mathrm{H}} \right]$. We now analyze the dynamic behaviour of the different PSDs. The early echo error PSD $\boldsymbol{\Psi}_\tau^{\mathrm{E_{early}}}$ is assumed to be time-variant and its norm decreases during convergence of the AF partitions $\hat{\boldsymbol{w}}_{b,\tau}$. In contrast, we can assume the PSD matrix of the late echo and background noise $\boldsymbol{\Psi}_\tau^{\mathrm{P}} = \boldsymbol{\Psi}_\tau^{\mathrm{D_{late}}} + \boldsymbol{\Psi}_\tau^{\mathrm{N}}$ to be only slowly time-varying. This is motivated by the temporal smoothing effect resulting from the tails of room impulse responses (RIRs) [17] and the characteristics of many background noise signals, e.g., microphone noise or babble noise. On the other hand, the near-end speaker PSD matrix $\boldsymbol{\Psi}_\tau^{\mathrm{S}}$ is modelled to be potentially rapidly time-varying following the dynamics of speech signals.

While all state-of-the-art approaches aim at a direct estimation of $\boldsymbol{\Psi}_\tau^{\mathrm{I}}$ from the prior error signal $\boldsymbol{e}_\tau^+$, we propose the additive observation noise PSD estimator

$$\hat{\boldsymbol{\Psi}}_\tau^{\mathrm{I}} = \hat{\boldsymbol{\Psi}}_\tau^{\mathrm{P}} + \hat{\boldsymbol{\Psi}}_\tau^{\mathrm{S}} \quad (20)$$

which allows to exploit the different time-variance of the statistics of $\boldsymbol{p}_\tau$ and $\boldsymbol{s}_\tau$. Considering PFs that are designed to extract the desired near-end speaker $\tilde{\boldsymbol{s}}_\tau$ with minimum distortion from the prior error $\tilde{\boldsymbol{e}}_\tau^+$ (cf. Eq. (16)), a straightforward estimator for the near-end PSD $\boldsymbol{\Psi}_\tau^{\mathrm{S}}$ is given by the periodogram of the masked prior error

$$\left[\hat{\boldsymbol{\Psi}}_\tau^{\mathrm{S}} \right]_{mm} = \lambda_S \left[\hat{\boldsymbol{\Psi}}_{\tau-1}^{\mathrm{S}} \right]_{mm} + (1-\lambda_S) \left| \left[\widehat{\boldsymbol{M}}_\tau \boldsymbol{e}_\tau^+ \right]_m \right|^2 \quad (21)$$

with the recursive averaging factor λ_S. Note that due to the same DFT length M the estimated STFT mask $\widehat{\boldsymbol{M}}_\tau$ can be similarly applied in the overlap-save domain. By subtracting the near-end PSD matrix estimate $\hat{\boldsymbol{\Psi}}_\tau^{\mathrm{S}}$ from Eq. (19) and assuming $\boldsymbol{\Psi}_\tau^{\mathrm{S}} \approx \hat{\boldsymbol{\Psi}}_\tau^{\mathrm{S}}$, the prior error PSD matrix is given by $\boldsymbol{\Psi}_\tau^{\mathrm{E}} \approx \boldsymbol{\Psi}_\tau^{\mathrm{E_{early}}} + \boldsymbol{\Psi}_\tau^{\mathrm{P}}$. As the late echo and background noise PSD matrix $\boldsymbol{\Psi}_\tau^{\mathrm{P}}$ varies only slowly compared to the early echo error PSD matrix $\boldsymbol{\Psi}_\tau^{\mathrm{E_{early}}}$, any stationary noise PSD estimator can be used for its inference. In this paper we use the minimum statistics estimator [18] which is given by the minimum of the latest κ estimates

$$\left[\hat{\boldsymbol{\Psi}}_\tau^{\mathrm{P}} \right]_{mm} = \min \left[\left[\boldsymbol{\Upsilon}_{\tau-\kappa+1}^{\mathrm{P}} \right]_{mm}, \quad \ldots, \quad \left[\boldsymbol{\Upsilon}_\tau^{\mathrm{P}} \right]_{mm} \right] \quad (22)$$

Algorithm 1 Proposed KF+DPF algorithm for one block of microphone samples.

Compute prior error $\boldsymbol{e}_\tau^+$ (cf. Eq. (13))
Infer DPF mask $\widehat{\boldsymbol{M}}_\tau$ (cf. Sec. III-B)
Update PSD estimates $\hat{\boldsymbol{\Psi}}_\tau^{\mathrm{I}}$, $\hat{\boldsymbol{\Psi}}_{b,\tau}^{\Delta\mathrm{W}}$ (cf. Eqs. (20) and (24))
Update AF estimates $\hat{\boldsymbol{w}}_{b,\tau}, \boldsymbol{P}_{b,\tau}$ (cf. Eq. (13))
Compute time-domain near-end signal $\underline{\hat{s}}_\tau$ by inverse STFT

of a smoothed periodogram

$$\left[\boldsymbol{\Upsilon}_\tau^{\mathrm{P}}\right]_{mm} = \lambda_P \left[\boldsymbol{\Upsilon}_{\tau-1}^{\mathrm{P}}\right]_{mm} + (1-\lambda_P)\left|[\hat{\boldsymbol{p}}_\tau]_m\right|^2 \tag{23}$$

with the late echo and background noise estimate $\hat{\boldsymbol{p}}_\tau = \left(\boldsymbol{I}_M - \widehat{\boldsymbol{M}}_\tau\right)\boldsymbol{e}_\tau^+$ and the recursive averaging factor λ_P. Note that $\hat{\boldsymbol{\Psi}}_\tau^{\mathrm{P}}$ can be interpreted as a temporally smoothly changing minimum regularization in the KF step size (13). Finally, the process noise PSD matrices are estimated by [7]

$$\left[\hat{\boldsymbol{\Psi}}_{b,\tau}^{\Delta\mathrm{W}}\right]_{mm} = (1-A^2)\ \left[\hat{\boldsymbol{\Psi}}_{b,\tau}^{\mathrm{W}}\right]_{mm} \tag{24}$$

with $\hat{\boldsymbol{\Psi}}_{b,\tau}^{\mathrm{W}} = \lambda_W \hat{\boldsymbol{\Psi}}_{b,\tau-1}^{\mathrm{W}} + (1-\lambda_W)\hat{\boldsymbol{w}}_{b,\tau-1}\hat{\boldsymbol{w}}_{b,\tau-1}^{\mathrm{H}}$.

V. Algorithmic Description

The proposed echo cancellation scheme for one block of microphone samples $\boldsymbol{y}_\tau$ is illustrated in Fig. 1 and summarized in Alg. 1. After computing the prior error $\boldsymbol{e}_\tau^+$, using the previous AF estimate (cf. Eq. (13)), the mask $\widehat{\boldsymbol{M}}_\tau$ is inferred by the DPF (cf. Sec. III-B). Subsequently, the PSD matrices $\boldsymbol{\Psi}_{b,\tau}^{\Delta\mathrm{W}}$ and $\boldsymbol{\Psi}_\tau^{\mathrm{I}}$ are estimated by Eqs. (24) and (20), respectively. Note that if the initial matrices are chosen to be diagonal, the estimators ensure all subsequent estimates to be diagonal as well. Afterwards, the means $\hat{\boldsymbol{w}}_{b,\tau}$ and state uncertainty matrices $\boldsymbol{P}_{b,\tau}$ of the AF partitions $\boldsymbol{w}_{b,\tau}$ are updated by the KF (cf. Eq. (13)) using the estimated PSD matrices $\hat{\boldsymbol{\Psi}}_\tau^{\mathrm{I}}$ and $\hat{\boldsymbol{\Psi}}_{b,\tau}^{\Delta\mathrm{W}}$. Finally, the time-domain near-end signal $\underline{\hat{s}}_\tau$ is computed by applying the inverse STFT to the DPF estimate $\hat{\underline{\boldsymbol{s}}}_\tau$.

VI. Experiments

In this section, the proposed algorithm is evaluated for a large variety of AEC scenarios. Each scenario is created by randomly drawing a far-end and a near-end speech signal from the *LibriSpeech* database [19] comprising 283 different speakers. Subsequently, the clean echo signal $\underline{\boldsymbol{d}}_\tau$ is simulated by convolving the far-end signal with a randomly drawn RIR from the databases [20], [21], [22], comprising 201 different RIRs with reverberation times T_{60} ranging from 120 ms to 780 ms. Finally, the near-end speaker signal and white Gaussian sensor noise are added. Both signals are scaled according to a random near-end-to-echo and echo-to-noise ratio in the ranges $[-10,\ 10]$ dB and $[30,\ 35]$ dB, respectively. An EPC is simulated by randomly drawing RIRs and signals for simulating the observations before and after a specific switching time. The switching time was chosen randomly in between 7.2 s and 8.8 s to avoid overfitting of the DNN to a fixed point in time. The block shift was set to $R = 256$ samples with a sampling frequency of $f_s = 16$ kHz. The PBKF modelled

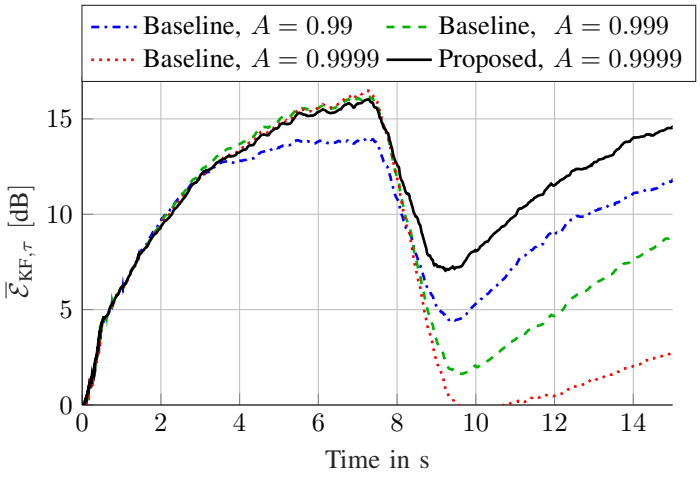

Fig. 2. Time-dependent ERLE of the PBKF for various parametrizations of the baseline noise PSD estimators and the proposed estimator.

$B = 8$ partitions which corresponds to a filter length of $L = 2048$ samples. The noise PSD estimators used the parameters $\lambda_P = \lambda_W = 0.9$, $\lambda_S = 0$ and $\kappa = 90$. The input features to the DPF were computed by using a Hamming window and $\epsilon_1 = 10^{-12}$. The DPF used approximately 3.4 million parameters with the input dimension of the stacked GRU layers being $P = 512$. It was trained using the ADAM optimizer [23] with a step size of 10^{-3}, a regularization factor of $\epsilon_2 = 10^{-12}$ and 4.4 hours of training data. The training data was preprocessed by a PBKF for which the required PSD matrices $\boldsymbol{\Psi}_\tau^{\mathrm{S}}$ and $\boldsymbol{\Psi}_{b,\tau}^{\Delta\mathrm{W}}$ were estimated by Eqs. (20) - (24). Here, the estimated mask $\widehat{\boldsymbol{M}}_\tau$ was replaced by an oracle mask. The testing data (27 mins) was disjoint from the training data, i.e., different speakers and RIRs.

In the first experiment the reconvergence behaviour of the PBKF after abrupt EPCs is compared for the proposed PSD estimator (cf. Sec. IV) relative to the state-of-the-art approach [24], i.e., recursively averaging the prior error power $|\,[\boldsymbol{e}_\tau^+]_m\,|^2$ with an averaging factor of 0.5. We compared several choices for the state transition parameter A because [8] reports a trade-off between steady-state performance and reconvergence behaviour. As performance measure we used the time-dependent logarithmic Echo Return Loss Enhancement (ERLE)

$$\mathcal{E}_{\mathrm{KF},\tau} = 10\log_{10}\frac{\mathbb{E}\left[||\underline{\boldsymbol{d}}_\tau||^2\right]}{\mathbb{E}\left[||\underline{\boldsymbol{d}}_\tau - \hat{\underline{\boldsymbol{d}}}_\tau||^2\right]}, \tag{25}$$

with $||\cdot||^2$ denoting the squared Euclidean norm and the expectation being approximated by temporal recursive averaging. To allow for more general conclusions, the time-dependent logarithmic ERLE $\mathcal{E}_{\mathrm{KF},\tau}$ has been averaged over 100 different scenarios. The resulting average time-dependent ERLE $\overline{\mathcal{E}}_{\mathrm{KF},\tau}$ for various choices of the state transition parameter A for the baseline and the proposed noise PSD estimator is shown in Fig. 2. As can be concluded from Fig. 2 for the baseline, larger state transition parameters A result in better steady-state performance at the cost of slower reconvergence after EPCs due to an overestimation of the noise PSD. The proposed noise

TABLE I
MEAN AND STANDARD DEVIATION (IN PARENTHESES) OF THE VARIOUS COMPONENTS OF THE PROPOSED AEC ALGORITHM. FOR THE BEST PERFORMANCE VALUES BOLD FONT IS USED.

	PBKF-only	DPF-only	PBKF+DPF
t_{pr}[ms] / RTF	**0.4**/**0.03**	1.0/0.06	1.4/0.09
$\overline{\mathcal{E}}_{(\cdot)}$	10.5 (2.7)	13.2 (2.8)	**17.0** (3.6)
$\overline{\mathcal{S}}_{\text{PF}}$	∞	14.6 (3.2)	**26.4** (5.7)
$\overline{\Delta\text{PESQ}}$	0.55 (0.4)	0.67 (0.4)	**1.12** (0.5)

PSD estimator, however, avoids this trade-off and allows for both high steady-state performance and rapid reconvergence.

Finally, we evaluate the echo suppression and near-end distortion performance of the individual algorithmic components, i.e., PBKF-only and DPF-only, and their synergistic combination PBKF+DPF. As performance measures we use:

$$\mathcal{E}_{\text{PF}} = 10\log_{10}\frac{||\underline{\boldsymbol{d}}||^2}{||\text{pf}(\underline{\boldsymbol{d}}-\widehat{\underline{\boldsymbol{d}}})||^2},\ \mathcal{S}_{\text{PF}} = 10\log_{10}\frac{||\beta\underline{\boldsymbol{s}}||^2}{||\beta\underline{\boldsymbol{s}}-\text{pf}(\underline{\boldsymbol{s}})||^2},$$

$$\mathcal{E}_{\text{KF}} = 10\log_{10}\frac{||\underline{\boldsymbol{d}}||^2}{||\underline{\boldsymbol{d}}-\widehat{\underline{\boldsymbol{d}}}||^2},\quad \Delta\text{PESQ} = \text{pq}(\underline{\boldsymbol{s}},\hat{\underline{\boldsymbol{s}}}) - \text{pq}(\underline{\boldsymbol{s}},\underline{\boldsymbol{y}}).$$

Here, the ERLE averaged over the entire signal duration, denoted by omitting the time index τ, after the PBKF and the DPF is represented by $\mathcal{E}_{\text{KF}}$ and $\mathcal{E}_{\text{PF}}$, respectively. The near-end distortion is measured by $\mathcal{S}_{\text{PF}}$, with the scaling factor $\beta = \frac{\boldsymbol{s}^{\text{T}}\text{pf}(\boldsymbol{s})}{||\boldsymbol{s}||^2}$ [25], and the PESQ (Perceptual Evaluation of Speech Quality [26]) improvement ΔPESQ. Note that the processing of a signal by the DPF is described by $\text{pf}(\cdot)$ while $\text{pq}(\cdot,\cdot)$ denotes the computation of the PESQ [26]. Tab. I shows the arithmetic averages of 100 random experiments of the performance measures $\mathcal{E}_{(\cdot)}$, $\mathcal{S}_{\text{PF}}$ and ΔPESQ, denoted by an overbar. Furthermore, the runtime t_{pr} to process one signal block on an *Intel Xeon CPU E3-1275 v6 @ 3.80GHz* and the corresponding Real-Time Factor (RTF) are given. Note that only the proposed noise PSD estimator (cf. Sec. IV) is evaluated due to the limited reconvergence capabilities of the baseline estimator. Furthermore, to show the effect of the PBKF, we evaluated a DPF-only algorithm which was trained with the microphone signal $\underline{\boldsymbol{y}}_\tau$ instead of the error signal $\underline{\boldsymbol{e}}^+_\tau$ in the feature computation (14). We conclude from Tab. I that while using only a PBKF results in limited echo cancellation, the DPF-only approach introduced significant distortions. In contrast the combination allows for high echo attenuation while introducing only little distortions. Finally, we see from Tab. I that due to the modest computational requirements, the proposed method is well suited for real-time applications.

VII. CONCLUSION

In this paper, we proposed a novel synergistic KF+DPF algorithm which improves state-of-the-art AEC algorithms for time-varying acoustic scenarios in terms of reconvergence speed without compromising performance for static scenarios. This is achieved by efficiently exploiting the DPF near-end estimate and the KF estimation error for inferring an AF step-size. Without any auxiliary mechanisms, we thereby overcome the limitations of KF-based step-size adaptation algorithms.

REFERENCES

[1] G. Enzner et al., "Acoustic Echo Control," in *Academic Press Library in Signal Processing*, vol. 4, pp. 807–877. Elsevier, FL, USA, 2014.

[2] B. Widrow and M. E. Hoff, "Adaptive Switching Circuits," in *Proc. WESCON Conv. Rec.*, Los Angeles, USA, Aug. 1960, pp. 96–104.

[3] E. Ferrara, "Fast implementations of LMS adaptive filters," *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 474–475, Aug. 1980.

[4] H. Zhang et al., "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions," in *Interspeech*, Sept. 2019, pp. 4255–4259.

[5] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network," 2020, arXiv: 2010.14337.

[6] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Process.*, vol. 86, no. 6, pp. 1140–1156, June 2006.

[7] F. Kuech et al., "State-space architecture of the partitioned-block-based acoustic echo controller," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 1295–1299.

[8] F. Yang et al., "Frequency-Domain Adaptive Kalman Filter With Fast Recovery of Abrupt Echo-Path Changes," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1778–1782, Dec. 2017.

[9] T. Jiang et al., "An Improved Practical State-Space FDAF With Fast Recovery of Abrupt Echo-Path Changes," *IEEE Access*, vol. 7, pp. 61353–61362, 2019.

[10] T. Haubner et al., "Noise-robust adaptation control for supervised acoustic system identification exploiting a noise dictionary," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021.

[11] G. Carbajal et al., "Multiple-Input Neural Network-Based Residual Echo Suppression," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Apr. 2018, pp. 231–235.

[12] L. Pfeifenberger and F. Pernkopf, "Nonlinear Residual Echo Suppression Using a Recurrent Neural Network," in *Interspeech*, Oct. 2020, pp. 3950–3954.

[13] Mhd. M. Halimeh et al., "Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021.

[14] G. Carbajal et al., "Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise," *IEEE Audio, Speech, and Language Process.*, vol. 28, pp. 2158–2173, 2020.

[15] Y. Xia et al., "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Barcelona, Spain, May 2020, pp. 871–875.

[16] A. A. Nugraha et al., "Multichannel Audio Source Separation With Deep Neural Networks," *IEEE Audio, Speech, and Language Process.*, vol. 24, no. 9, pp. 1652–1664, Sept. 2016.

[17] H. Kuttruff, *Room Acoustics, Sixth Edition*, CRC Press, 2016.

[18] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[19] V. Panayotov et al., "Librispeech: An ASR corpus based on public domain audio books," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[20] M. Jeub et al., "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Int. Conf. on Digit. Signal Process.*, Santorini, Greece, July 2009.

[21] J. Y. C. Wen et al., "Evaluation of speech dereverberation algorithms using the mardy database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2006.

[22] "Multi-channel impulse response database (MIRD)," https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/multi-channel-impulse-response-database, Accessed: 2020-12-04.

[23] D. Kingma and J. Ba, "ADAM: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] J. Franzen and T. Fingscheidt, "Improved Measurement Noise Covariance Estimation for N-channel Feedback Cancellation Based on the Frequency Domain Adaptive Kalman Filter," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Brighton, UK, May 2019, pp. 965–969.

[25] J. LeRoux et al., "SDR – Half-baked or Well Done?," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Brighton, UK, May 2019, pp. 626–630.

[26] ITU-T Recommendation P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Recommendation, ITU, Nov. 2007.