Joint Online Estimation of Early and Late Residual Echo PSD for Residual Echo Suppression

Naveen Kumar Desiraju, Simon Doclo, *Senior Member, IEEE*, Markus Buck, *Member, IEEE*, and Tobias Wolff

***Abstract*—In hands-free telephony and other distant-talking applications, an acoustic echo cancellation system is typically required, where a short adaptive filter is often used in practice to achieve fast convergence at low computational cost. This may result in late residual echo (LRE) remaining due to under-modeling of the echo path and early residual echo (ERE) due to filter misalignment. Both residual echo components can be suppressed using a postfilter in the subband domain, which requires accurate estimates of the power spectral density (PSD) of the ERE and LRE components. State-of-the-art methods estimate the ERE and LRE PSDs independently of each other, where the ERE PSD is estimated by simply multiplying the loudspeaker PSD with a frequency-dependent scalar and the LRE PSD is estimated using a recursive estimator based on frequency-dependent reverberation scaling and decay parameters. In this paper, we propose to extend the ERE PSD estimator from a scalar to a moving average filter on the loudspeaker PSD. In addition, we propose a signal-based method to jointly estimate all model parameters for the ERE and LRE PSD estimators in online mode, and derive two gradient-descent-based algorithms to simultaneously update the model parameters by minimizing the mean squared log error. The proposed method is compared with state-of-the-art methods in terms of estimation accuracy of the model parameters as well as the residual echo PSDs. Simulation results using both artificially generated as well as measured impulse responses show that the proposed method outperforms state-of-the-art methods for all considered scenarios.**

***Index Terms*—Acoustic echo cancellation, adaptive filters, PSD estimation, residual echo suppression.**

I. Introduction

HANDS-FREE telephony and distant-talking applications have become very popular in recent years. In these applications, the distance between the desired (near-end) speaker and the microphone may be quite large, while the loudspeaker playing back the far-end signal is typically located much closer to the microphone. As a result, the microphone signal may be degraded significantly due to the acoustic echo of the far-end signal, which may lead to the near-end speaker being unintelligible. In a typical acoustic echo cancellation (AEC) system, an adaptive filter aims at estimating the impulse response (IR) between the loudspeaker and the microphone [1], [2]. In practice, however, this filter is typically not able to perfectly estimate the IR, resulting in residual echo. In addition, as a short filter is often used in practice to achieve fast convergence at low computational cost, the filter is unable to estimate the complete echo path. Thus, assuming no non-linear signal components, the residual echo is composed of early residual echo (ERE) due to filter misalignment and late residual echo (LRE) due to under-modeling of the IR by the short AEC filter.

The residual echo is often suppressed in the subband domain using a postfilter, for which both model-based approaches [3], [4], [5], [6], [7], [8], [9], [10] as well as deep learning-based approaches [11], [12], [13], [14] have been proposed. In this paper, we focus on model-based approaches to estimate the power spectral density (PSD) of both the ERE and LRE components. A simple but frequently used method is to estimate the ERE PSD as a scaled version of the PSD of either the far-end signal [1] or the estimated echo signal generated by the AEC filter [15]. In either case, the scalar (referred to as coupling factor) is estimated during periods of near-end speech absence as the ratio between the PSD of the AEC error signal and the PSD of the respective input signal. To estimate the LRE PSD, several methods have been proposed based on the statistical reverberation model in [16], which assumes that the late reverberant part of an IR decays exponentially at a rate proportional to the reverberation time. A recursive estimator for the LRE PSD was proposed in [6], which requires estimates of two frequency-independent room acoustic parameters: the reverberation scaling parameter (related to the initial power of the LRE component) and the reverberation decay parameter (related to the reverberation time). Both reverberation parameters were estimated using a *channel-based* method, i.e., using the coefficients of the converged AEC filter. In [7], a similar recursive estimator for the LRE PSD was derived with frequency-dependent reverberation parameters, where both parameters were again estimated using a channel-based method. Since channel-based methods are effective only if the AEC filter is long enough to capture a significant portion of the decay of the IR, *signal-based* methods have also have been proposed, where the reverberation parameters are estimated using the

Manuscript received 11 January 2022; revised 30 July 2022 and 29 October 2022; accepted 6 November 2022. Date of publication 23 November 2022; date of current version 2 December 2022. This work was supported in part by the European Union's Seventh Framework Programme under Grant FP7/2007-2013, and in part by the Project DREAMS under Grant ITN-GA-2012-316969. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Keisuke Kinoshita. *(Corresponding author: Naveen Kumar Desiraju.)*

Naveen Kumar Desiraju was with the Acoustic Speech Enhancement Research, Nuance Communications Deutschland GmbH, 89077 Ulm, Germany. He is now with the Department of Audio and Media Technologies, Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany (e-mail: naveen.kumar.desiraju@iis.fraunhofer.de).

Simon Doclo is with the Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4all, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: simon.doclo@uni-oldenburg.de).

Markus Buck and Tobias Wolff are with the Cerence Inc., 89077 Ulm, Germany (e-mail: markus.buck@cerence.com; tobias.wolff@cerence.com).

Digital Object Identifier 10.1109/TASLP.2022.3224292

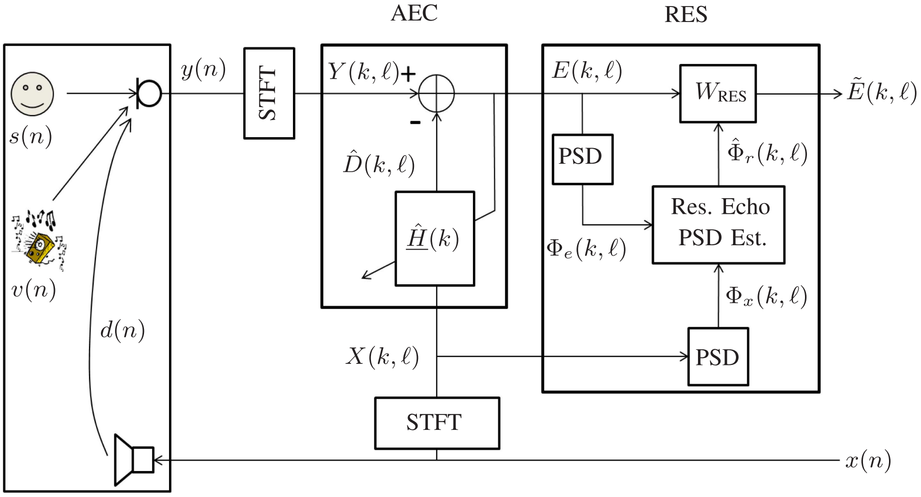

Fig. 1. Acoustic echo cancellation (AEC) and residual echo suppression (RES) systems.

far-end and residual echo signals. A recursive estimator for the LRE PSD was derived in [8] based on the generalized reverberation model in [17], where a signal-based method was proposed to estimate the reverberation parameters in offline mode (i.e., batch processing). In [9], we proposed signal-based methods to *jointly* estimate both reverberation parameters in online mode by minimizing either the mean squared error (MSE) or the mean squared log error (MSLE) cost function. In [10], a coupling-factor-based estimator for the early acoustic echo PSD and a recursive estimator for the late acoustic echo PSD were considered in a pure acoustic echo suppression system (i.e., without an AEC filter). A signal-based method exploiting higher-order statistics was proposed to estimate the parameters independently of each other in online mode.

As an extension of [9], in this paper we propose a signal-based method to estimate the PSD of both residual echo components based on parametric models. First, we propose to model the ERE PSD using a moving average filter (instead of a coupling factor) on the PSD of the far-end signal. By assuming that the filter misalignment is spread evenly over all AEC filter taps in the subband domain, we can model all coefficients of the moving average filter using a single (frequency-dependent) parameter. Similarly as in [7], [9], the LRE PSD is modeled using an IIR filter on the PSD of the far-end signal based on (frequency-dependent) reverberation scaling and decay parameters. Second, we propose to jointly estimate all three model parameters in online mode using the output error method [18], [19], [20], [21] by minimizing an MSLE cost function. To simultaneously update the model parameters, we use gradient-descent-based algorithms such as recursive prediction error and pseudo-linear regression, which were originally derived for time-domain recursive systems [19]. The proposed method is first evaluated in an idealistic setting, i.e., using artificially generated IRs and no AEC filter. It is then compared with state-of-the-art model-based methods [1] and [10] in a realistic setting, i.e., using measured IRs and a slowly converging AEC filter, in terms of estimation accuracy of the residual echo PSD and the resulting residual echo suppression and near-end speech distortion.

The remainder of the paper is organized as follows. The signal model as well as the AEC and postfilter systems are introduced in Section II. In Section III, the considered models for the ERE and LRE PSDs are presented. In Section IV, we discuss state-of-the-art methods for estimating the different model parameters and present the proposed method for jointly estimating the model parameters in online mode. In Section V, simulation results using artificial as well as measured IRs are presented.

II. Signal Model, AEC and Postfilter Systems

Fig. 1 shows a loudspeaker-enclosure-microphone system in which the far-end signal x is played through the loudspeaker and the microphone captures the acoustic echo d, the near-end speech s and the background noise v. The microphone signal at discrete-time sample n is thus given as:

$$y(n) = s(n) + v(n) + \underbrace{\sum_{i=0}^{N_h-1} h(i) \cdot x(n-i)}_{d(n)}, \qquad (1)$$

where h denotes the IR between the loudspeaker and the microphone, which is assumed to be time-invariant and of length N_h samples. To remove the acoustic echo from the microphone signal, we consider a system in the subband domain consisting of two parts [1]: an AEC filter and a residual echo suppression (RES) filter.

A. Acoustic Echo Cancellation

We consider a G-tap subband AEC filter $\underline{\hat{H}}$, with the filter length G chosen so as to cover only the direct path and early reflections in h. For subband processing, the (windowed) time-domain signals are transformed into the short-time Fourier transform (STFT) domain using a fast Fourier transform (FFT) filterbank of order N_{FFT}, with the total number of subbands $K = \frac{N_{\mathrm{FFT}}}{2} + 1$. The complex-valued spectrum of the far-end signal x in the k^{th} subband and ℓ^{th} frame is given as:

$$X(k,\ell) = \sum_{i=0}^{N_{\text{FFT}}-1} x(\ell \cdot F + i) \cdot W_{\text{ana}}(i) \cdot e^{-j\frac{2\pi}{N_{\text{FFT}}}ki}, \tag{2}$$

where $j = \sqrt{-1}$, F denotes the frameshift and W_{ana} denotes the analysis window. The spectra of the other time-domain signals are computed similarly to (2), with the spectral equivalent of (1) given as:

$$Y(k,\ell) = S(k,\ell) + V(k,\ell) + D(k,\ell). \tag{3}$$

The acoustic echo estimate is generated by filtering the far-end signal through the AEC filter:

$$\hat{D}(k,\ell) = \underline{X}^H(k,\ell)\ \underline{\hat{H}}(k), \tag{4}$$

where $\underline{X}(k,\ell) = [X(k,\ell) \quad \ldots \quad X(k,\ell-G+1)]^T$ denotes the G-dimensional input vector to the subband AEC filter $\underline{\hat{H}}$, $(\cdot)^H$ denotes the Hermitian operator and $(\cdot)^T$ denotes the transpose operator. The AEC error signal is then given as:

$$\begin{aligned} E(k,\ell) &= Y(k,\ell) - \hat{D}(k,\ell) \\ &= S(k,\ell) + V(k,\ell) + \Big(D(k,\ell) - \hat{D}(k,\ell)\Big) \\ &= S(k,\ell) + V(k,\ell) + R(k,\ell) \\ &= S(k,\ell) + V(k,\ell) + \underbrace{R_E(k,\ell)}_{\text{Misalignment}} + \underbrace{R_L(k,\ell)}_{\text{Under-modeling}}, \end{aligned} \tag{5}$$

where R, R_E and R_L denote the residual echo, ERE and LRE components, respectively. The ERE component is given as:

$$R_E(k,\ell) = \underline{X}^H(k,\ell)\ \Delta\underline{H}_E(k), \tag{6}$$

with the AEC misalignment filter defined as:

$$\Delta\underline{H}_E(k) = \underline{H}_E(k) - \underline{\hat{H}}(k), \tag{7}$$

where $\underline{H}_E$ contains the first G coefficients of the equivalent subband filter corresponding to h. Since in this paper $G = \lfloor \frac{N}{F} \rfloor$, where $N \ll N_h$ corresponds to the length of the direct path and early reflections in h, the LRE component R_L is assumed to contain only late reflections, also known as reverberation.

B. Residual Echo Suppression

From (5), it can be observed that in addition to the desired near-end speech signal, the AEC error signal also contains background noise and residual echo components. It is desirable to suppress these interfering components while maintaining high quality and low distortion of the near-end speech signal. As shown in Fig. 1, this suppression is performed by applying a postfilter W_{RES} to the AEC error signal E. A frequently used postfilter is the Wiener gain [1], i.e.:

$$W_{\text{RES}}(k,\ell) = 1 - \left(\frac{\lambda_r(k,\ell) + \lambda_v(k,\ell)}{\lambda_e(k,\ell)}\right), \tag{8}$$

where λ_r, λ_v and λ_e denote the PSDs of the residual echo, background noise and AEC error signals, respectively. Assuming that S, V and R are mutually uncorrelated, the PSD of the AEC error signal can be expressed using (5) as:

$$\lambda_e(k,\ell) = \mathcal{E}\left\{|E(k,\ell)|^2\right\} = \lambda_s(k,\ell) + \lambda_v(k,\ell) + \lambda_r(k,\ell), \tag{9}$$

where $\mathcal{E}\{\cdot\}$ denotes the statistical expectation operator. Additionally, we assume that the early and late residual echo components are uncorrelated, such that the residual echo PSD can be written as:

$$\lambda_r(k,\ell) = \lambda_{r_E}(k,\ell) + \lambda_{r_L}(k,\ell), \tag{10}$$

where λ_{r_E} and λ_{r_L} denote the ERE PSD and LRE PSD, respectively.

In practice, the statistical expectation operator in (9) can be approximated by temporal averaging of the periodogram, i.e.:

$$\Phi_e(k,\ell) = \alpha \cdot \Phi_e(k,\ell-1) + (1-\alpha) \cdot |E(k,\ell)|^2, \tag{11}$$

where Φ_e is an approximation of λ_e and α denotes the recursive smoothing factor. For an unobservable signal such as the residual echo R, the quantity Φ_r itself needs to be estimated, with its estimate denoted as $\hat{\Phi}_r$. In the remainder of this paper, we will use the term *PSD* to refer to the quantities λ and Φ and the term *PSD estimate* to refer to $\hat{\Phi}$.

In order to control the aggressiveness of the residual echo suppression, we will use the following gain [1]:

$$\boxed{W_{\text{RES}}(k,\ell) = \max\left\{1 - \beta \cdot \left(\frac{\hat{\Phi}_r(k,\ell) + \hat{\Phi}_v(k,\ell)}{\Phi_e(k,\ell)}\right), \gamma\right\}}, \tag{12}$$

where β denotes the over-estimation factor and γ denotes the spectral floor, i.e., the maximum attenuation of the filter. Based on (10), the residual echo PSD estimate is given by:

$$\boxed{\hat{\Phi}_r(k,\ell) = \hat{\Phi}_{r_E}(k,\ell) + \hat{\Phi}_{r_L}(k,\ell).} \tag{13}$$

Although during near-end speech and noise absence, $\hat{\Phi}_r$ can be easily estimated from Φ_e based on (9), this is obviously not possible during periods of double-talk. Hence, in this paper we will use parametric models for the ERE PSD Φ_{r_E} and the LRE PSD Φ_{r_L}, which will be explained in the next section. Many approaches have been proposed in literature to estimate the PSD of the background noise $\hat{\Phi}_v$ [22], [23], [24]. In this paper, we assume that the background noise is stationary and its PSD estimate $\hat{\Phi}_v$ is known.

The processed AEC error signal is given as:

$$\tilde{E}(k,\ell) = W_{\text{RES}}(k,\ell) \cdot E(k,\ell), \tag{14}$$

which can be expressed as the sum of its individual components similarly to (5):

$$\tilde{E}(k,\ell) = \tilde{S}(k,\ell) + \tilde{V}(k,\ell) + \tilde{R}(k,\ell), \tag{15}$$

where $\tilde{S}$, $\tilde{V}$ and $\tilde{R}$ are obtained in simulations by independently filtering S, V and R with W_{RES}. The processed signals $\tilde{E}$, $\tilde{S}$ and $\tilde{R}$ are synthesized into the time-domain using inverse STFT and overlap-add processing to yield the processed time-domain signals $\tilde{e}$, $\tilde{s}$ and $\tilde{r}$, respectively. These signals can then be used to evaluate the near-end speech distortion and residual echo suppression (see Section V-C).

III. Models for Early and Late Residual Echo PSD

In this section, we present the considered parametric models for the early and late residual echo PSDs. We propose to model the ERE PSD using a moving average filter on the PSD of the far-end signal. Similarly as in [7], [9], the LRE PSD is modeled using an IIR filter on the PSD of the far-end signal.

A. Model for Early Residual Echo PSD

As already mentioned, the ERE is caused by the misalignment between the AEC filter and the IR. A simple model for the ERE PSD was proposed in [1], where the ERE PSD is a scaled version of the PSD of the far-end signal:

$$\hat{\Phi}_{r_E}(k,\ell) = C(k) \cdot \Phi_x(k,\ell), \tag{16}$$

where C denotes the (frequency-dependent) coupling factor. As shown in [1], the coupling factor represents the squared magnitude spectrum of the filter misalignment. A disadvantage of this model is that a scalar coupling factor may not be sufficient to model the ERE PSD, especially if a long AEC filter is used.

We now derive our proposed model for the ERE PSD. Using (6), the ERE PSD is given as:

$$\begin{aligned}\lambda_{r_E}(k,\ell) &= \mathcal{E}\left\{|R_E(k,\ell)|^2\right\}\\ &= \mathcal{E}\left\{\left|\sum_{g=0}^{G-1} X^*(k,\ell-g)\cdot\Delta H_E(k,g)\right|^2\right\},\\ &= \mathcal{E}\Bigg\{\sum_{i=0}^{G-1}\sum_{j=0}^{G-1} X^*(k,\ell-i)\cdot X(k,\ell-j)\\ &\qquad \cdot\Delta H_E(k,i)\cdot\Delta H_E^*(k,j)\Bigg\},\end{aligned} \tag{17}$$

where $\Delta H_E(k,g)$ denotes the g^{th} coefficient of the AEC misalignment filter $\Delta\underline{H}_E$ in (7). Assuming statistical independence between the far-end signal and the AEC misalignment filter yields:

$$\begin{aligned}\lambda_{r_E}(k,\ell) = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} &\mathcal{E}\left\{X^*(k,\ell-i)\cdot X(k,\ell-j)\right\}\\ &\cdot\,\mathcal{E}\left\{\Delta H_E(k,i)\cdot\Delta H_E^*(k,j)\right\}.\end{aligned} \tag{18}$$

Assuming that the coefficients of the AEC misalignment filter are mutually uncorrelated, i.e., $\mathcal{E}\{\Delta H_E(k,i)\cdot\Delta H_E^*(k,j)\} = 0$ for $i \neq j$, the ERE PSD can be written as:

$$\lambda_{r_E}(k,\ell) = \sum_{g=0}^{G-1}\lambda_x(k,\ell-g)\cdot\mathcal{E}\left\{|\Delta H_E(k,g)|^2\right\}. \tag{19}$$

Finally, assuming that the misalignment is spread evenly over all AEC filter coefficients[1] [1], [25], [26], i.e., $\mathcal{E}\{|\Delta H_E(k,g)|^2\} = C(k)\ \forall g$, the ERE PSD can be simplified as:

$$\lambda_{r_E}(k,\ell) = C(k)\cdot\sum_{g=0}^{G-1}\lambda_x(k,\ell-g). \tag{20}$$

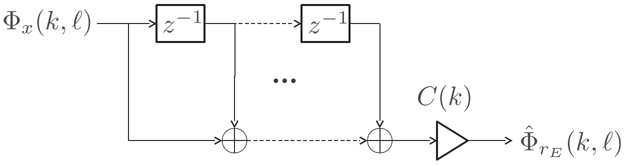

Fig. 2. Proposed model for the ERE PSD Φ_{r_E} (moving average filter).

Based on (20), we will hence use the following model for the ERE PSD:

$$\boxed{\hat{\Phi}_{r_E}(k,\ell) = C(k)\cdot\sum_{g=0}^{G-1}\Phi_x(k,\ell-g).} \tag{21}$$

This model can be interpreted as an extension of (16) in that a moving average filter is used instead of an instantaneous scaling of the PSD of the far-end signal. This model is depicted in Fig. 2.

B. Model for Late Residual Echo PSD

As already mentioned, the LRE component is caused by under-modeling of the IR by the AEC filter. Several models for the LRE PSD have been proposed based on the statistical reverberation model in [16], which assumes that the late reverberant part of an IR can be described as an exponentially decaying realization of a stochastic process:

$$h(i) = w_L(i)\cdot e^{-\rho\,(i-N)}, \qquad N \leq i < N_h, \tag{22}$$

where $w_L \sim \mathcal{N}(0,\sigma_L^2)$ is a zero-mean white Gaussian noise process with variance σ_L^2, ρ denotes the decay rate and i denotes the filter coefficient index. The decay rate is related to the reverberation time T_{60} of the room as:

$$\rho = \frac{3\cdot\ln 10}{f_s\cdot T_{60}}, \tag{23}$$

where f_s denotes the sampling rate in Hz. It should be noted that in practice T_{60}, and hence the decay rate ρ, are frequency-dependent [27]. Based on (22), a recursive expression for the LRE PSD was first derived in [6] using frequency-independent parameters. In this paper, we will use a version of this model with frequency-dependent parameters, which was derived in [7] and [9], and is given as:

$$\boxed{\hat{\Phi}_{r_L}(k,\ell) = A(k)\cdot\Phi_x(k,\ell-G) + B(k)\cdot\hat{\Phi}_{r_L}(k,\ell-1),} \tag{24}$$

where $A(k)$ and $B(k)$ denote the frequency-dependent reverberation scaling and decay parameters, respectively. These parameters are related to the frequency-dependent variance $\sigma_L^2(k)$ and decay rate $\rho(k)$ as [9]:

$$A(k) = \sigma_L^2(k)\cdot\left(\frac{1-e^{-2\rho(k)F}}{1-e^{-2\rho(k)}}\right), \tag{25}$$

$$B(k) = e^{-2\rho(k)F}. \tag{26}$$

[1] It should be noted that this assumption may be violated if the AEC filter coefficients diverge, which can happen if there are echo path changes or if there are errors in the detection of double-talk periods.

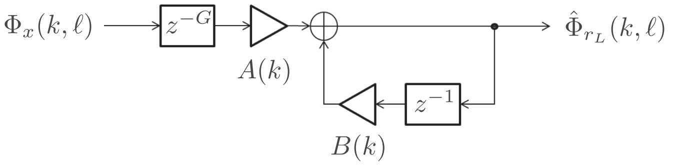

Fig. 3. Model for the LRE PSD Φ_{r_L} (IIR filter).

The recursive expression in (24) is depicted in Fig. 3 as an IIR filter on the PSD of the far-end signal.

IV. Parameter Estimation Methods

In Section IV-A, we briefly review state-of-the-art signal-based methods to estimate the model parameters A, B and C. In Section IV-B, we present our proposed signal-based methods to jointly estimate all model parameters by minimizing a single cost function. Please note that in all considered methods, the model parameters are estimated only during periods of near-end speech absence.

A. State-of-the-art Methods

In this section, we briefly discuss state-of-the-art methods for estimating the three model parameters A, B and C.

Hänsler et al. [1] estimate the scalar coupling factor C in (16) as the smoothed ratio of the AEC error PSD and the PSD of the far-end signal:

$$\hat{C}_{\mathrm{H}}(k,\ell) = (1-\delta)\cdot\frac{\Phi_e(k,\ell)}{\Phi_x(k,\ell)} + \delta\cdot\hat{C}_{\mathrm{H}}(k,\ell-1), \tag{27}$$

where δ denotes a smoothing factor. Please note that in [1], no additional estimator for the LRE PSD was used, i.e., the estimated coupling factor from (27) was fed into (16) to yield an estimate for the total residual echo PSD $\hat{\Phi}_r$.

Favrot et al. [10] considered a pure acoustic echo suppression setup (i.e., no AEC filter) and proposed a coupling-factor-based estimator for the early acoustic echo PSD as well as a recursive estimator for the late acoustic echo PSD. The model parameters were estimated independently of each other in online mode using a method based on higher-order statistics. In order to facilitate a fair comparison, in this paper we consider a modified version of Favrot's method as a benchmark to estimate all three model parameters, and therefore both the ERE and LRE PSDs, in the presence of an AEC filter (see Appendix A).

B. Joint Parameter Estimation Method

Based on the parametric models for the ERE and LRE PSDs in (21) and (24), in this section we propose a method to *jointly* estimate all three model parameters A, B and C in *online* mode. This method is an extension of the method in [9], which assumed no filter misalignment ($\Phi_{r_E}=0$), and therefore only estimated the reverberation parameters A and B. To jointly estimate the parameters of generic IIR filters in the time-domain, several signal-based methods have been proposed [18], [19], [20], [21], either based on output error (OE) or equation error (EE). In [9] we investigated both the OE and EE methods (applied to PSDs) to jointly estimate the reverberation parameters A and B, either using the MSE or MSLE cost function. Simulation results showed that the OE method using the MSLE cost function yielded the best performance in terms of PSD estimation accuracy and residual echo suppression. Therefore, in this paper we will only consider the OE method using the MSLE cost function to jointly estimate all model parameters (reverberation parameters A and B and coupling factor C).

By merging the moving average filter model for the ERE PSD in (21) with the recursive model for the LRE PSD in (24), the residual echo PSD estimate in (13) is given as:

$$\boxed{\begin{aligned}\hat{\Phi}_r(k,\ell) = {}& \hat{C}(k,\ell)\cdot\sum_{g=0}^{G-1}\Phi_x(k,\ell-g)\\ &+\hat{A}(k,\ell)\cdot\Phi_x(k,\ell-G)+\hat{B}(k,\ell)\cdot\hat{\Phi}_{r_L}(k,\ell-1),\end{aligned}} \tag{28}$$

where $\hat{A}(k,\ell)$, $\hat{B}(k,\ell)$ and $\hat{C}(k,\ell)$ denote estimates of the model parameters in subband k and frame ℓ and can be represented by the vector:

$$\underline{\hat{\theta}}(k,\ell) = \left[\begin{array}{ccc}\hat{A}(k,\ell) & \hat{B}(k,\ell) & \hat{C}(k,\ell)\end{array}\right]^T. \tag{29}$$

From (28), it can be observed that the PSD estimate in the current frame $\hat{\Phi}_r(k,\ell)$ not only depends on the parameter estimates in the current frame $\underline{\hat{\theta}}(k,\ell)$ but also on the PSD estimate $\hat{\Phi}_r(k,\ell-1)$, which itself depends on the parameter estimates in the previous frame $\underline{\hat{\theta}}(k,\ell-1)$, and so on. Thus, $\hat{\Phi}_r$ is a non-linear function of $\underline{\hat{\theta}}$, where the current PSD estimate depends on parameter estimates in all previous frames.

To update all model parameters in each frame, we consider the instantaneous MSLE cost function:

$$J\left(\ln\hat{A}(k,\ell),\,\ln\hat{B}(k,\ell),\,\ln\hat{C}(k,\ell)\right) = Q_{\ln}^2(k,\ell), \tag{30}$$

where $Q_{\ln}$ denotes the logarithmic error between the target PSD Φ_r and the PSD estimate $\hat{\Phi}_r$ in (28):

$$Q_{\ln}(k,\ell) = \ln\left(\frac{\Phi_r(k,\ell)}{\hat{\Phi}_r(k,\ell)}\right). \tag{31}$$

Similarly as in [9], we now derive gradient-descent-based algorithms to update the model parameters $\underline{\hat{\theta}}(k,\ell)$. Since the residual echo PSD Φ_r is obviously not observable, we will only update the model parameters during periods of near-end speech absence and when the AEC error signal is not dominated by background noise, such that we can replace Φ_r by Φ_e in (31). The estimated model parameters will then be used, both during periods of near-end speech absence as well as double-talk, to estimate the residual echo PSD $\hat{\Phi}_r$. The block scheme to estimate the model parameters in online mode is depicted in Fig. 4. Please note that even though good noise reduction algorithms are available, any errors in the estimation of the background noise PSD and/or detection of double-talk periods can have a negative impact on the accuracy of the parameter estimates.

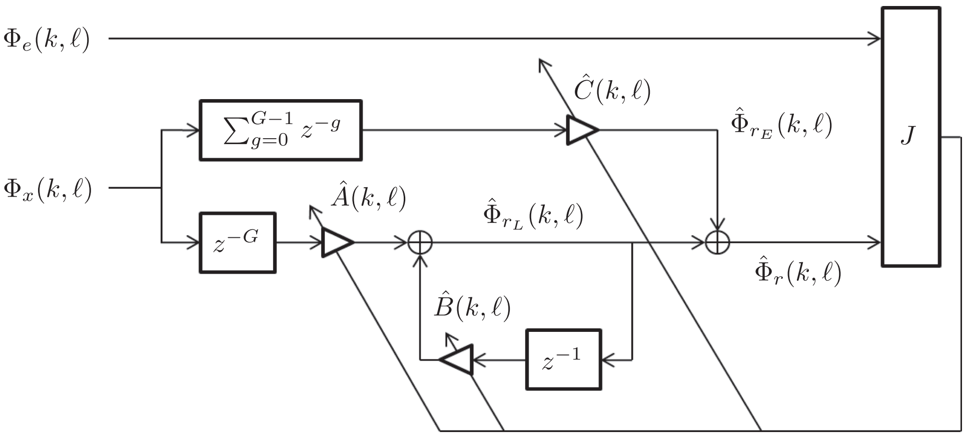

Fig. 4. Online joint estimation of the three model parameters using the output error method by minimizing a single cost function.

The gradient-descent update rule in the logarithmic domain is given as:

$$\boxed{\ln \hat{\theta}(k,\ell+1) = \ln \hat{\theta}(k,\ell) - \frac{\mu_\theta}{2} \cdot J'_\theta(k,\ell),} \tag{32}$$

where $\theta \in \{A, B, C\}$ denotes a model parameter and μ_θ denotes the step-size used to update it. J'_θ denotes the partial derivative of the cost function J in (30) w.r.t. the logarithm of the parameter estimate $\ln \hat{\theta}$, and is computed using (30) and (31) as:

$$\begin{aligned} J'_\theta(k,\ell) &= \frac{\partial Q^2_{\ln}(k,\ell)}{\partial \ln \hat{\theta}(k,\ell)} = 2 \cdot Q_{\ln}(k,\ell) \cdot \frac{\partial Q_{\ln}(k,\ell)}{\partial \ln \hat{\theta}(k,\ell)} \\ &= -2 \cdot \frac{Q_{\ln}(k,\ell)}{\hat{\Phi}_r(k,\ell)} \cdot \frac{\partial \hat{\Phi}_r(k,\ell)}{\partial \ln \hat{\theta}(k,\ell)}. \end{aligned} \tag{33}$$

Using (28), the partial derivative $\frac{\partial \hat{\Phi}_r(k,\ell)}{\partial \ln \hat{\theta}(k,\ell)}$ for the three model parameters is equal to:

$$\begin{aligned} \frac{\partial \hat{\Phi}_r(k,\ell)}{\partial \ln \hat{A}(k,\ell)} = &\hat{A}(k,\ell) \cdot \Phi_x(k,\ell-G) + \\ &\hat{B}(k,\ell) \cdot \frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{A}(k,\ell)}, \end{aligned} \tag{34}$$

$$\begin{aligned} \frac{\partial \hat{\Phi}_r(k,\ell)}{\partial \ln \hat{B}(k,\ell)} = &\hat{B}(k,\ell) \cdot \hat{\Phi}_{r_L}(k,\ell-1) + \\ &\hat{B}(k,\ell) \cdot \frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{B}(k,\ell)}, \end{aligned} \tag{35}$$

$$\frac{\partial \hat{\Phi}_r(k,\ell)}{\partial \ln \hat{C}(k,\ell)} = \hat{C}(k,\ell) \cdot \sum_{g=0}^{G-1} \Phi_x(k,\ell-g). \tag{36}$$

It can be observed that the right hand side of (34) and (35) contain the partial derivatives of the LRE PSD estimate in the *previous* frame $\hat{\Phi}_{r_L}(k,\ell-1)$ w.r.t. the logarithms of the parameter estimates in the *current* frame $\ln \hat{A}(k,\ell)$ and $\ln \hat{B}(k,\ell)$, respectively. These terms exist due to the recursive model for the LRE PSD in (24). These partial derivatives cannot be computed in a straightforward manner, as $\hat{\Phi}_{r_L}(k,\ell-1)$ does not directly depend on either $\hat{A}(k,\ell)$ or $\hat{B}(k,\ell)$. In the following, we present two algorithms which have been proposed in [19] to approximate these partial derivatives.

1) Recursive Prediction Error (RPE): The RPE algorithm approximates the partial derivatives using the parameter estimates in the previous frame:

$$\frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{A}(k,\ell)} \approx \frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{A}(k,\ell-1)},$$

$$\frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{B}(k,\ell)} \approx \frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{B}(k,\ell-1)}, \tag{37}$$

which are reasonable approximations if the step-sizes μ_A and μ_B used to update the reverberation parameters are sufficiently small. Using (37) in (34) and (35) enables to compute the partial derivatives recursively. This method will be referred to as OE-RPE-MSLE.

2) Pseudo Linear Regression (PLR): The PLR algorithm simply assumes that the LRE PSD estimate in the previous frame $\hat{\Phi}_{r_L}(k,\ell-1)$ is independent of the parameter estimates in the current frame, i.e.:

$$\frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{A}(k,\ell)} = 0, \qquad \frac{\partial \hat{\Phi}_{r_L}(k,\ell-1)}{\partial \ln \hat{B}(k,\ell)} = 0. \tag{38}$$

It should be noted that these assumptions are stronger than for the RPE algorithm in (37). Using (38) in (34) and (35) yields non-recursive formulations for the partial derivatives, which are therefore approximate versions of the partial derivatives computed using the RPE algorithm. This method will be referred to as OE-PLR-MSLE.

V. Simulation Results

In this section, we evaluate the performance of the proposed parameter estimation methods, i.e., OE-RPE-MSLE and OE-PLR-MSLE, and compare their performance with the state-of-the-art signal-based methods discussed in Section IV-A. We will refer to the proposed methods estimating all three model parameters as 3P methods, whereas we will refer to the simplified versions in [9] estimating only the two reverberation parameters as 2P methods.

In Sections V-A and V-B, we present the acoustic conditions and the algorithmic parameters used in the simulations. In Section V-C we discuss the performance metrics used to evaluate the PSD estimation accuracy, residual echo suppression and near-end speech distortion. In Section V-D, we present the simulation results for two settings: an idealistic setting using artificially generated IRs and a realistic setting using real-world IRs. Please note that an analysis of the performance of the proposed parameter estimation methods in the presence of different types and amount of noise and/or non-linear echo components, even though highly relevant to the overall performance evaluation of the proposed methods, is out-of-scope of this paper.

A. Acoustic Conditions

For all simulations, the sampling frequency of the time-domain signals is equal to $f_s = 16$ kHz. For the far-end signal x, we have considered five speech sequences (3 male and 2 female) of length 10 s each, while for the near-end signal s, we have considered five speech sequences (2 male and 3 female) of length 5 s each, resulting in 25 different combinations of far-end and near-end speakers. All speech sequences have been chosen from the TIMIT database [28], and for each combination of far-end and near-end speaker, the double-talk condition occurs in the last 5 s. The 10 s long background noise signal v is stationary air conditioner noise measured in a quiet office.

Two different types of IRs have been considered in the simulations:

- Artificial IRs: the artificial IRs have been generated according to the following time-domain model:

$$\Delta h(i) = \begin{cases} w_E(i), & 0 \leq i < N \\ w_L(i) \cdot e^{-\rho\,(i-N)}, & N \leq i < N_h, \end{cases} \tag{39}$$

  where $w_E \sim \mathcal{N}(0, \sigma_E^2)$ and $w_L \sim \mathcal{N}(0, \sigma_L^2)$ are zero-mean white Gaussian noise processes with variances σ_E^2 and σ_L^2, respectively, and ρ denotes the decay rate defined in (23). This model assumes that the first N coefficients of Δh correspond to the AEC misalignment filter (in the time-domain), where the misalignment is spread evenly over all AEC filter coefficients, whereas the latter coefficients of Δh correspond to the exponentially decaying model in (22). The IR parameters σ_L^2 and ρ are related to the (frequency-independent) model parameters A and B as in (25) and (26), while the IR parameter σ_E^2 is related to the (frequency-independent) parameter C as $C = \sigma_E^2 \cdot F$ (see Appendix B). A total of 180 artificial IRs have been generated using all combinations of the frequency-independent parameters σ_E^2, σ_L^2 and T_{60} given in Table I, with $N = 640$ and $N_h = 16000$. These values represent a wide-ranging and realistic set of acoustic conditions, with the values of σ_E^2 representing low to high amounts of early residual echo (due to different amounts of echo cancellation), and the values of σ_L^2 and T_{60} representing dry to highly reverberant rooms.

TABLE I
PARAMETER VALUES FOR GENERATING THE ARTIFICIAL IRS

Parameter	Values
σ_E^2	$\{-60, -50, -40, -30, -20, -10\}$ dB
σ_L^2	$\{-40, -36, -32, -28, -24, -20\}$ dB
T_{60}	$\{200, 400, 600, 800, 1000\}$ ms

TABLE II
DETAILS ABOUT MEASURED IRS

Room	No. of IRs	$\mathbf{T_{60}}$	Shape
Lab	16	300-400 ms	Rectangular
Garage	16	400-500 ms	Rectangular
Office	16	500-600 ms	L-shaped
Echoic	7	850-950 ms	Rectangular

TABLE III
STEP-SIZES USED FOR THE PROPOSED METHODS

Method	$\mu_{\mathbf{A}}$	$\mu_{\mathbf{B}}$	$\mu_{\mathbf{C}}$
OE-RPE-MSLE	$10^{-1.5}$	10^{-4}	$10^{-1.5}$
OE-PLR-MSLE	$10^{-1.5}$	10^{-3}	$10^{-1.5}$

- Measured IRs: A total of 55 IRs have been measured in four rooms with different reverberation times, with details given in Table II. The broadband T_{60} for each IR has been estimated via line-fitting on its corresponding energy decay curve [29].

B. Algorithmic Parameters

For the subband processing, a filterbank of order $N_{\text{FFT}} = 512$ (i.e., $K = 257$) and an overlap of 75% (i.e., frameshift $F = 128$) have been used, with a Hann window as the analysis window. To estimate the PSDs in (11) from the periodograms, we have used a recursive smoothing factor $\alpha = e^{\frac{-2 \cdot F}{f_s \cdot t_c}}$ with the time-constant $t_c = 0.02$s corresponding to the typically assumed stationarity of speech signals.[2] For the postfilter in (12), an over-estimation factor $\beta = 2$ and a spectral floor $\gamma = -20$ dB have been used. Based on the model for the ERE PSD in (21), we choose the filter length of the moving average filter in (28) to be the same as the AEC filter length $G\ (= 5)$.[3] For the proposed methods, the step-sizes listed in Table III were determined via a brute-force search procedure such that each parameter estimation method yielded its most optimal performance. For the state-of-the-art methods, the following parameters have been used:

[2] It should be noted that using no recursive smoothing ($\alpha = 0$) or using only recursive smoothing corresponding to G frames ($\alpha = \frac{G-1}{G+1}$) without the moving average filter in (21) yields worse results.

[3] It should be noted that using a shorter or longer filter length for the moving average filter results in under- or over-modeling of the underlying system, respectively, resulting in performance deterioration.

- Hänsler's method [1]: smoothing factor $\delta = 0.9$ in (27)
- Modified Favrot's method (see Appendix A): $N = 640$, $O = 1024$ and $P = \kappa \cdot F$ with $\kappa = 12$.

C. Performance Metrics

To evaluate the estimation accuracy of the residual echo PSD, we consider the Log Spectral Distance (LSD) [24] between the target PSD Φ_r and the residual echo PSD estimate $\hat{\Phi}_r$ in (28), defined as:

$$\mathrm{LSD} = \frac{10}{K \cdot L_1} \cdot \sum_{k=0}^{K-1} \sum_{\ell=l_1+1}^{l_1+L_1} \left| \log_{10} \left(\frac{\Phi_r(k,\ell)}{\hat{\Phi}_r(k,\ell)} \right) \right|, \quad (40)$$

where l_1 and L_1 denote the start and the duration of the evaluation window in frames, respectively. We choose the evaluation window to be between 4 s and 5 s (before double-talk starts), i.e., $l_1 = 500$ and $L_1 = 125$. If the LSD score is low, it means that the residual echo PSD estimate is accurate, with the perfect estimate $\hat{\Phi}_r(k,\ell) = \Phi_r(k,\ell)$ resulting in $\mathrm{LSD} = 0$.

To evaluate the amount of residual echo suppression after applying the postfilter, we consider the segmental residual echo attenuation [9], defined as:

$$\mathrm{REA}_{\mathrm{seg}} = \frac{10}{L_1} \cdot \sum_{\ell=l_1+1}^{l_1+L_1} \log_{10} \left(\frac{\sum_{f=0}^{F-1} r^2(\ell \cdot F + f)}{\sum_{f=0}^{F-1} \tilde{r}^2(\ell \cdot F + f)} \right), \quad (41)$$

where the time-domain signals r and $\tilde{r}$ are obtained through inverse STFT processing of the residual echo signal R and its postfiltered version $\tilde{R}$, respectively (see Section II-B). If the $\mathrm{REA}_{\mathrm{seg}}$ score is high, it means that a large amount of residual echo has been suppressed, which is desirable.

Similarly, to evaluate the amount of near-end speech distortion, we consider the segmental speech-to-speech distortion ratio [30], defined as:

$$\mathrm{SSDR}_{\mathrm{seg}} = \frac{10}{L_2} \cdot \sum_{\ell=l_2+1}^{l_2+L_2} \log_{10} \left(\frac{\sum_{f=0}^{F-1} s^2(\ell \cdot F + f)}{\sum_{f=0}^{F-1} s_d^2(\ell \cdot F + f)} \right), \quad (42)$$

where $s_d(n) = s(n) - \tilde{s}(n)$, with $\tilde{s}$ obtained through inverse STFT processing of the postfiltered near-end speech signal $\tilde{S}$. This score is computed during periods of double-talk, which occurs between 5 s and 10 s, i.e., $l_2 = 625$ and $L_2 = 625$. If the $\mathrm{SSDR}_{\mathrm{seg}}$ score is high, it means that the distortion of the near-end speech signal is low, which is desirable.

D. Experimental Results

The first experiment is performed in an idealistic setting using artificially generated IRs, no near-end speech and no background noise. This experiment aims at evaluating the estimation accuracy of the proposed 3P methods, the simplified 2P versions in [9] as well as Favrot's method for the artificial IR parameters and the residual echo PSD. The second experiment is performed in a realistic setting using measured IRs, near-end speech, background noise and a slowly converging subband AEC filter. This experiment aims at comparing the PSD estimation accuracy and the residual echo suppression performance between the proposed 3P methods, the simplified 2P versions and the considered state-of-the-art methods.

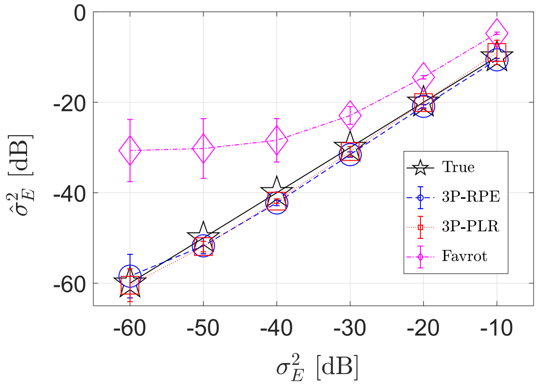

Fig. 5. Variance of the misalignment: $\hat{\sigma}_E^2$ vs. σ_E^2 for the proposed methods and Favrot's method in the idealistic setting.

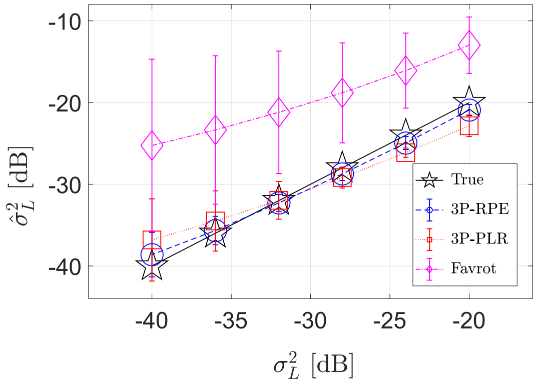

Fig. 6. Variance of the late part of the IR: $\hat{\sigma}_L^2$ vs. σ_L^2 for the proposed methods and Favrot's method in the idealistic setting.

1) Idealistic Setting: In this experiment, we use the artificially generated IRs (see Table I) to generate the acoustic echo signal and do not consider an AEC filter, i.e., $\underline{\hat{H}}(k) = [0 \dots 0]^T$. Additionally, we assume no near-end speech ($s = 0$) and background noise ($v = 0$). This means that the microphone signal in (1) is given as:

$$y(n) = d(n) = \sum_{i=0}^{N_h-1} \Delta h(i) \cdot x(n-i), \quad (43)$$

with Δh defined in (39) and $E(k,\ell) = Y(k,\ell)$. For this idealistic setting, we evaluate the accuracy of the residual echo PSD estimate $\hat{\Phi}_r$ obtained using the proposed 3P methods, the simplified 2P versions as well as Favrot's method, and compare the estimates of the artificial IR parameters $\hat{\sigma}_E^2$, $\hat{\sigma}_L^2$ and $\hat{T}_{60}$ with the true values. These parameter estimates are obtained by averaging the converged values of $\hat{A}(k)$, $\hat{B}(k)$ and $\hat{C}(k)$ over all frequency bins and feeding them in (25), (26), and (52), respectively.

Figs. 5, 6, and 7 show the true variance of the misalignment σ_E^2 against the estimated variance $\hat{\sigma}_E^2$, the true variance of the late part of the IR σ_L^2 against the estimated variance $\hat{\sigma}_L^2$, and the true reverberation time T_{60} against the estimated reverberation time $\hat{T}_{60}$ for the proposed 3P methods as well as Favrot's method. Each point in Fig. 5 is obtained by averaging

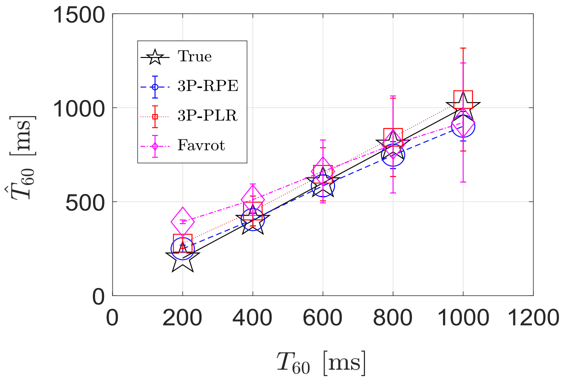

Fig. 7. Reverberation time: $\hat{T}_{60}$ vs. T_{60} for the proposed methods and Favrot's method in the idealistic setting.

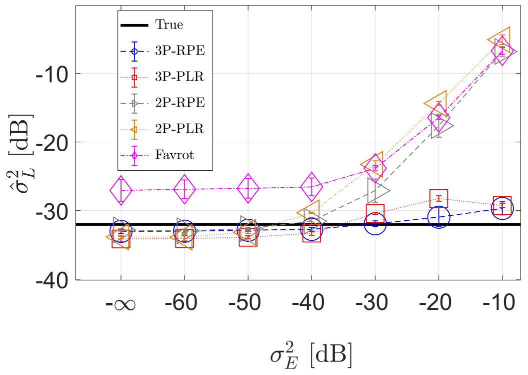

Fig. 8. Estimated variance $\hat{\sigma}_L^2$ obtained using the proposed methods (2P and 3P versions) and Favrot's method as a function of σ_E^2 in the idealistic setting ($\sigma_L^2 = -32$ dB).

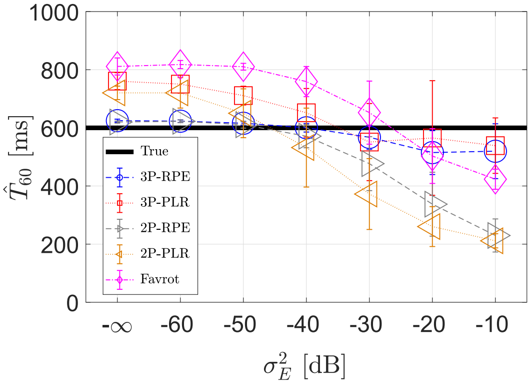

Fig. 9. Estimated reverberation time $\hat{T}_{60}$ obtained using the proposed methods (2P and 3P versions) and Favrot's method as a function of σ_E^2 in the idealistic setting ($T_{60} = 600$ ms).

the estimates $\hat{\sigma}_E^2$ over 30 IRs with different σ_L^2 and T_{60} values, each point in Fig. 6 is obtained by averaging the estimates $\hat{\sigma}_L^2$ over 30 IRs with different σ_E^2 and T_{60} values and each point in Fig. 7 is obtained by averaging the estimates $\hat{T}_{60}$ over 36 IRs with different σ_E^2 and σ_L^2 values. The error bars depict the standard deviations across the respective IRs. It can be observed from Fig. 5 that for both proposed 3P methods, the parameter σ_E^2 can be estimated very accurately (with very small standard deviations) over a large range of parameter values, indicating robustness to different values of σ_L^2 and T_{60}, while Favrot's method consistently over-estimates the parameter σ_E^2, with larger standard deviations. In addition, it can be observed from Figs. 6 and 7 that the RPE algorithm typically yields more accurate estimates (and especially smaller standard deviations) of the parameters σ_L^2 and T_{60} than the PLR algorithm over a large range of parameter values. This is not surprising, since the PLR algorithm is an approximation of the RPE algorithm. Favrot's method significantly over-estimates the parameter σ_L^2 over a large range of parameter values (with very large standard deviations), while it gives reasonably accurate estimates for the T_{60} parameter, albeit with large standard deviations.

We now investigate the benefit of estimating all three model parameters (3P) against estimating only two model parameters (2P) in the simplified versions presented in [9]. To this end, we compare the influence of different amounts of misalignment, represented by σ_E^2, on the estimation accuracy of the parameters σ_L^2 and T_{60}. For $\sigma_L^2 = -32$ dB, Fig. 8 shows the estimated variance $\hat{\sigma}_L^2$ obtained using the 2P and 3P estimation methods as well as Favrot's method for different values of σ_E^2. Each point is obtained by averaging the estimates over 6 IRs with different T_{60} values. For $T_{60} = 600$ ms, Fig. 9 shows the estimated reverberation time $\hat{T}_{60}$ obtained using the 2P and 3P estimation methods as well as Favrot's method for different values of σ_E^2. Each point is obtained by averaging the estimates over 6 IRs with different σ_L^2 values. It should be noted that $\sigma_E^2 = -\infty$ dB corresponds to no filter misalignment, i.e., no early residual echo. It can be observed that the 2P methods yield accurate estimates for σ_L^2 and T_{60} only for low values of σ_E^2, and fail to do so for large amounts of filter misalignment. On the other hand, the proposed 3P methods yield reasonably accurate estimates for σ_L^2 and T_{60} for all considered σ_E^2 values, where the RPE algorithm again outperforms the PLR algorithm. In contrast, Favrot's method consistently over-estimates σ_L^2, with the over-estimation increasing significantly for large amounts of filter misalignment, while it over-estimates T_{60} for low amounts of misalignment and under-estimates it for high amounts of misalignment, respectively. These results clearly show the benefit of estimating all three model parameters when using the proposed methods, especially when a significant amount of filter misalignment is present.

Fig. 10 shows the LSD scores between the target and the estimated residual echo PSDs, obtained using the 2P and 3P estimation methods as well as Favrot's method for different values of σ_E^2. Each point is obtained by averaging the LSD scores over 30 IRs with different σ_L^2 and T_{60} values, while the error bars depict the standard deviation across these IRs. It can be observed that the proposed 3P methods yield the most accurate estimates for the residual echo PSD, especially for large values of σ_E^2, with the RPE and PLR algorithms yielding similar results. In contrast, Favrot's method consistently yields the highest LSD scores irrespective of the amount of misalignment. This result again clearly shows the benefit of estimating all three model parameters when using the proposed methods.

2) Realistic Setting: In this experiment, we use IRs measured in different rooms (see Table II) to generate the acoustic echo

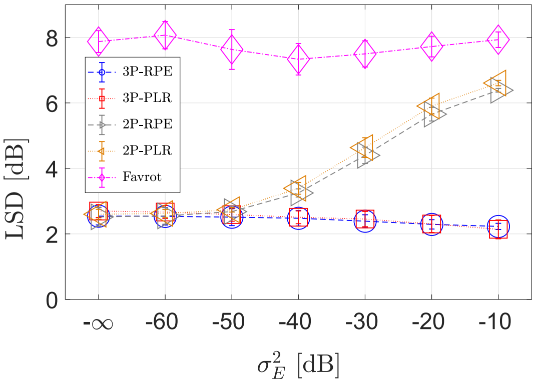

Fig. 10. LSD scores obtained using the proposed methods (2P and 3P versions) and Favrot's method as a function of σ_E^2 in the idealistic setting.

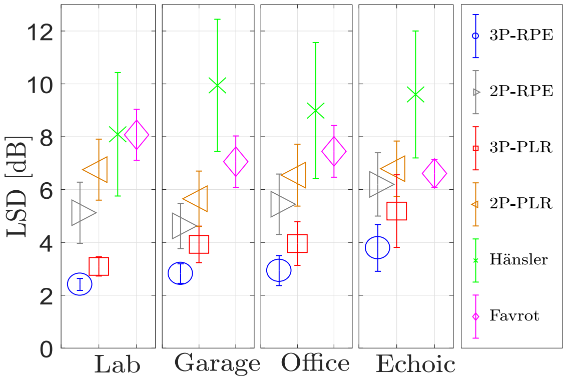

Fig. 11. LSD scores obtained using all considered parameter estimation methods for different rooms.

signal and we consider a slowly converging subband AEC filter $\underline{\hat{H}}$, whose coefficients are updated using the subband NLMS algorithm [21] with a small (fixed) stepsize of 5×10^{-3}. The length of the AEC filter is rather short ($G = 5$, corresponding to 64 ms), covering just the direct path and some early reflections in the IRs. Near-end speech is present at a signal-to-echo ratio of 0 dB, while background noise is present at a signal-to-noise ratio of 40 dB. As already mentioned, the model parameters are estimated only during periods of near-end speech absence, i.e., during the first 5 s, and only if the AEC error PSD Φ_e is at least 3 dB above the background noise PSD Φ_v. For this realistic setting, we compare the LSD, REA_{seg} and SSDR_{seg} scores between the proposed 3P methods, the simplified 2P versions and the considered state-of-the-art methods.

Fig. 11 shows the LSD scores obtained using all considered parameter estimation methods for different rooms. Each point is obtained by averaging the LSD scores over all IRs in a room, with the error bars depicting the standard deviation across these IRs. The rooms have been placed in order of increasing T_{60} from left to right. It can be observed that the proposed 3P method with the RPE algorithm consistently estimates the residual echo PSD more accurately than all other methods, with the next best performances delivered by the proposed 3P method with the PLR algorithm and the two simplified 2P versions. Hänsler's method, which uses just a single parameter (coupling factor) to estimate the complete residual echo PSD, yields the highest LSD scores for all rooms, which is to be expected, while Favrot's method delivers better results with increasing T_{60}.

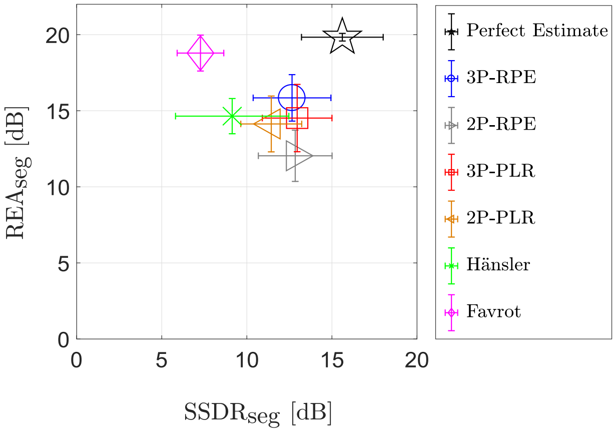

Fig. 12. Segmental residual echo attenuation (REA_{seg}) vs. segmental speech-to-speech distortion ratio (SSDR_{seg}) scores obtained using all considered parameter estimation methods for different rooms.

Fig. 12 shows the REA_{seg} scores plotted against the SSDR_{seg} scores obtained using all considered methods for all rooms. Each point is obtained by averaging the segmental metrics across IRs from all rooms, with the error bars on the x-axis and y-axis depicting the standard deviations across these IRs. For comparison, the scores obtained using the perfect residual echo PSD estimate $\hat{\Phi}_r(k,\ell) = \Phi_r(k,\ell)$ and an over-estimation factor $\beta = 1$ in (12) are also included, which corresponds to the best possible performance in terms of maximizing both segmental metrics. It can be observed that both proposed 3P methods as well as both simplified 2P versions yield the highest SSDR_{seg} scores (about 5-7 dB better than the other considered methods), with the 3P methods outperforming the 2P versions in terms of the REA_{seg} score (about 1–3 dB). In addition, it can be observed that Favrot's method yields the highest REA_{seg} score, but the proposed 3P method with the RPE algorithm clearly outperforms Favrot's method in terms of the SSDR_{seg} score. In conclusion, the proposed 3P method with the RPE algorithm provides the best performance in terms of maximizing both segmental metrics.

VI. Conclusion

In this paper, we assumed the residual echo to consist of only linear signal components and proposed two signal-based methods to jointly estimate the PSDs of the ERE and LRE components based on parametric models. We proposed to model the ERE PSD (due to filter misalignment) using a moving average filter on the PSD of the far-end signal, where we assumed that all coefficients of the moving average filter are the same (coupling factor). The LRE PSD (due to under-modeling of the echo path by the AEC filter) was modeled using a frequently used IIR filter on the PSD of the far-end signal, described by reverberation scaling and decay parameters. We proposed to jointly estimate all three model parameters in online mode using the output error method by minimizing an MSLE cost function, where the parameters are updated simultaneously using either the RPE or PLR algorithm. Simulation results using artificially generated IRs showed that the proposed methods yielded accurate estimates for the model parameters and the residual echo PSD, with the RPE algorithm performing better than the PLR algorithm. In addition, simulation results showed that jointly estimating all three model parameters is beneficial compared to only estimating the reverberation scaling and decay parameters, especially for high amounts of filter misalignment. Finally, simulation results using measured IRs showed that the proposed method with the RPE algorithm consistently outperformed all other methods in terms of PSD estimation accuracy and delivered the best performance in terms of maximizing the amount of residual echo suppression while minimizing the amount of near-end speech distortion.

Appendix A
Original and Modified Versions of Favrot's Method

In the original method in [10], the coupling factor C was estimated as:

$$\hat{C}_{\mathrm{F}}(k,\ell) = \frac{\mathcal{E}\left\{\tilde{\Phi}_y(k,\ell)\cdot\tilde{\Phi}_{x_M}(k,\ell)\right\}}{\mathcal{E}\left\{\tilde{\Phi}_{x_M}(k,\ell)\cdot\tilde{\Phi}_{x_M}(k,\ell)\right\}} = Z_M^y(k,\ell), \tag{44}$$

where $\tilde{\Phi}_{x_M}(k,\ell) = |X_M(k,\ell)|^2 - \Phi_{x_M}(k,\ell)$ and $\tilde{\Phi}_y(k,\ell) = |Y(k,\ell)|^2 - \Phi_y(k,\ell)$ represent the temporal fluctuations of the PSDs of the M-sample delayed far-end signal $x_M(n) = x(n-M)$ and the microphone signal $y(n)$, respectively. Here, Z_M^y is used to denote the ratio in (44) computed using the signals x_M and y. The delay $M(\ll N)$ was chosen so as to align the far-end signal x with the microphone signal y, i.e., it corresponds to the initial peak in the IR, which depends on the distance between the loudspeaker and the microphone. The decay rate B was estimated using (44) for two different delays O and $O+P$:

$$\hat{B}_{\mathrm{F}}(k,\ell) = \left(\frac{Z_{O+P}^y(k,\ell)}{Z_O^y(k,\ell)}\right)^{1/\kappa}, \tag{45}$$

where O corresponds to the late echo tail ($O \geq N$) and $P = \kappa \cdot F$ corresponds to a delay of κ frames.

The modification considered in this paper uses the temporal fluctuations of the AEC error PSD $\tilde{\Phi}_e(k,\ell)$ instead of $\tilde{\Phi}_y(k,\ell)$ to estimate the parameters B and C:

$$\begin{aligned}\hat{C}_{\mathrm{F}}(k,\ell) &= Z_M^e(k,\ell)\\ \hat{B}_{\mathrm{F}}(k,\ell) &= \left(\frac{Z_{O+P}^e(k,\ell)}{Z_O^e(k,\ell)}\right)^{1/\kappa}.\end{aligned} \tag{46}$$

Additionally, to estimate the parameter A, we use the N-sample delayed far-end signal x_N:

$$\hat{A}_{\mathrm{F}}(k,\ell) = Z_N^e(k,\ell). \tag{47}$$

Appendix B
Coupling Factor

Since in the time-domain the ERE signal $r_E(n)$ is equal to:

$$r_E(n) = \sum_{i=0}^{N-1} \Delta h(i)\cdot x(n-i), \tag{48}$$

its auto-correlation for lag τ is given as:

$$\begin{aligned}a_{r_E r_E}(n, n+\tau) &= \mathcal{E}\{r_E(n)\cdot r_E(n+\tau)\}\\ &= \mathcal{E}\left\{\sum_{i=0}^{N-1}\Delta h(i)\cdot x(n-i)\cdot\sum_{j=0}^{N-1}\Delta h(j)\cdot x(n-j+\tau)\right\}\\ &= \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\mathcal{E}\{\Delta h(i)\cdot\Delta h(j)\}\cdot a_{xx}(n-i, n-j+\tau),\end{aligned} \tag{49}$$

where a_{xx} denotes the auto-correlation of x. Using (39) and assuming that the far-end signal x is stationary over a short period of F samples, with $F \ll N$ $(= G\cdot F)$, we can rewrite (49) as:

$$\begin{aligned}a_{r_E r_E}(n, n+\tau) &= \sigma_E^2\cdot\sum_{i=0}^{N-1} a_{xx}(n-i, n-i+\tau),\\ &= \sigma_E^2\cdot\sum_{g=0}^{G-1}\sum_{f=0}^{F-1} a_{xx}(n-g\cdot F-f, n-g\cdot F-f+\tau)\\ &\approx \sigma_E^2\cdot F\cdot\sum_{g=0}^{G-1} a_{xx}(n-g\cdot F, n-g\cdot F+\tau).\end{aligned} \tag{50}$$

Applying the Wiener-Khinchin theorem to (50) yields:

$$\lambda_{r_E}(k,\ell) = \sigma_E^2\cdot F\cdot\sum_{g=0}^{G-1}\lambda_x(k,\ell-g), \tag{51}$$

such that comparing (51) with (20) yields:

$$C = \sigma_E^2\cdot F. \tag{52}$$

References

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. New York, NY, USA: Wiley, 2004.

[2] C. Breining et al., "Acoustic echo control - An application of very-high-order adaptive filters," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 42–69, Jul. 1999.

[3] C. Beaugeant, V. Turbin, P. Scalart, and A. Gillore, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Process.*, vol. 64, no. 1, pp. 33–47, 1998.

[4] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Process.*, vol. 10, no. 5, pp. 245–256, Jul. 2002.

[5] J. Franzen and T. Fingscheidt, "An efficient residual echo suppression for multi-channel acoustic echo cancellation based on the frequency-domain adaptive Kalman filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 226–230.

[6] G. Enzner, "A model-based optimum filtering approach to acoustic echo control: Theory and practice," Ph.D. dissertation, RWTH Aachen Univ., Aachen, Germany, 2006.

[7] E. Habets, I. Cohen, S. Gannot, and P. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1433–1451, Nov. 2008.

[8] M. Valero, E. Mabande, and E. Habets, "Signal-based late residual echo spectral variance estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 5914–5918.

[9] N. Desiraju, S. Doclo, M. Buck, and T. Wolff, "Online estimation of reverberation parameters for late residual echo suppression," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 77–91, 2020.

[10] A. Favrot, C. Faller, and F. Küch, "Modeling late reverberation in acoustic echo suppression," in *Proc. Int. Workshop Acoustic Signal Enhancement*, 2012, pp. 1–4.

[11] C. Lee, J. Shin, and N. Kim, “DNN-based residual echo suppression,” in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1775–1779.

[12] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, “Multiple-input neural network-based residual echo suppression,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 231–235.

[13] M. Halimeh, T. Haubner, A. Breigleb, A. Schmidt, and W. Kellermann, “Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 121–125.

[14] A. Ivry, I. Choen, and B. Berdugo, “Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 126–130.

[15] I. Schalk-Schupp, F. Faubel, M. Buck, and A. Wendemuth, “Combined linear and nonlinear residual echo suppression using a deficient distortion model — A proof of concept,” in *Proc. IEEE ITG Symp. Speech Commun.*, 2016, pp. 1–5.

[16] J. Polack, “La transmission de l’énergie sonore dans les salles,” Dissertation, Université du Maine, Le Mans, France, 1988.

[17] E. Habets, S. Gannot, and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[18] T. Söderström and P. Stoica, “Some properties of the output error method,” *Automatica*, vol. 18, no. 1, pp. 93–99, 1982.

[19] J. Shynk, “Adaptive IIR filtering,” *IEEE ASSP Mag.*, vol. 6, no. 2, pp. 4–21, Apr. 1989.

[20] Y. Tomita, A. Damen, and P. V. D. Hof, “Equation error versus output error methods,” *Ergonomics*, vol. 35, no. 5/6, pp. 551–564, 1992.

[21] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ, USA: Prentice Hall, 1996.

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

[23] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[24] T. Gerkmann and R. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[25] S. Yamamoto and S. Kitayama, “An adaptive echo canceller with variable step gain method,” *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. 65, pp. 1–8, 1982.

[26] F. Lindstrom, C. Schüldt, and I. Claesson, “An improvement of the two-path algorithm transfer logic for acoustic echo cancellation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1320–1326, May 2007.

[27] H. Kuttruff, *Room Acoustics*. London, U.K.: Spon Press, 2000.

[28] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, “DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM,” National Inst. of Standards and Technol., Gaithersburg, MD, USA, Internal Rep. 4930, 1990.

[29] M. Schroeder, “New method of measuring reverberation time,” *J. Acoust. Soc. Amer.*, vol. 37, pp. 409–412, 1965.

[30] T. Fingscheidt and S. Suhadi, “Quality assessment of speech enhancement systems by separation of enhanced speech, noise, and echo,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 818–821.

Naveen Kumar Desiraju received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Roorkee, India, in 2011, and the M.Sc. degree in signal processing and machine intelligence as a Commonwealth Scholar from the University of Surrey, Guildford, UK, in 2012. He is currently working toward the Doctoral degree with the Signal Processing Group, Department of Medical Physics and Acoustics, Carl von Ossietzky University, Oldenburg, Germany. From 2013 to 2017, he was a Doctoral Researcher with the Acoustic Speech Enhancement Team, Nuance Communications Deutschland GmbH, Ulm, Germany, on a Marie Skłodowska-Curie Fellowship. From 2018 to 2020, he was an Associate Engineer with the Automatic Speech Recognition Team, Harman Connected Services GmbH, Garching bei München, Germany. Since 2021, he has been employed as a Research Engineer with the Department of Audio and Media Techinolgies, Fraunhofer-Institut für Integrierte Schaltungen IIS, Erlangen, Germany. His research interests include speech enhancement, adaptive filtering, and machine learning.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering, and the Ph.D. degree in applied sciences from Katholieke Universiteit Leuven (KU), Leuven, Belgium, in 1997 and 2003. From 2003 to 2007 he was a Postdoctoral Fellow with KU Leuven and McMaster University, Hamilton, ON, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors in Leuven, Belgium. Since 2009, he has been a full Professor with the University of Oldenburg, Oldenburg, Germany, and a Scientific Advisor for the Oldenburg Branch for Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology IDMT. His research interests include signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. Prof. Doclo was the recipient of several best paper awards (International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019). He is Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and the EAA Technical Committee on Audio Signal Processing. Since 2021 he has been a Senior Area Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing.

Markus Buck (Member, IEEE) received the Dipl.-Ing. degree in electrical engineering and the Ph.D. degree from Ulm University, Ulm, Germany, in 1998 and 2004, respectively. From 1998 to 2009, he was with Temic Speech Dialog Systems and Harman/Becker Automotive Systems, Ulm, Germany, working on various topics in the field of speech processing. From 2009 to 2019, he had the role of a Research Manager with Nuance Communications, Ulm, Germany, leading the technology development in acoustic speech enhancement for hands-free telephony, speech recognition, and in-car communication. Since 2019, he has been with Cerence Inc., Ulm, Germany, continuing in the same role. His main research interests include multi-channel signal processing, adaptive filtering and neural network based methods for speech signal processing.

Tobias Wolff received the Dipl.-Ing. degree and the Dr.-Ing. degree in electrical engineering and communications from Signal Processing Group, Technische Universität Darmstadt, Darmstadt, Germany, in 2006 and 2011, respectively. Between 2005 and 2007, he was a Visiting Researcher with the Image Processing Laboratory, University of California Santa Barbara, Santa Barbara, CA, USA, working on subjective perception of video coding artifacts. In April 2009, he joined the Department of Speech Signal Enhancement, Nuance Communications Deutschland GmbH, Ulm, Germany. Since 2017, he has been a Principal Researcher with Nuance in the area of multimicrophone acoustic speech enhancement. Since 2019, he has been with Cerence Inc. in the same domain. His main scientific research interests include beamforming, source separation and acoustic localization of sound sources.