Integrated Sidelobe Cancellation and Linear Prediction Kalman Filter for Joint Multi-Microphone Speech Dereverberation, Interfering Speech Cancellation, and Noise Reduction

Thomas Dietzen, Simon Doclo, *Senior Member, IEEE*, Marc Moonen, *Fellow, IEEE*, and Toon van Waterschoot, *Member, IEEE*

***Abstract*—In multi-microphone speech enhancement, reverberation as well as additive noise and/or interfering speech are commonly suppressed by deconvolution and spatial filtering, e.g., using multi-channel linear prediction (MCLP) on the one hand and beamforming, e.g., a generalized sidelobe canceler (GSC), on the other hand. In this article, we consider several reverberant speech components, whereof some are to be dereverberated and others to be canceled, as well as a diffuse (e.g., babble) noise component to be suppressed. In order to perform both deconvolution and spatial filtering, we integrate MCLP and the GSC into a novel architecture referred to as integrated sidelobe cancellation and linear prediction (ISCLP), where the sidelobe-cancellation (SC) filter and the linear prediction (LP) filter operate in parallel, but on different microphone signal frames. Within ISCLP, we estimate both filters jointly by means of a single Kalman filter. We further propose a spectral Wiener gain post-processor, which is shown to relate to the Kalman filter's posterior state estimate. The presented ISCLP Kalman filter is benchmarked against two state-of-the-art approaches, namely first a pair of alternating Kalman filters respectively performing dereverberation and noise reduction, and second an MCLP+GSC Kalman filter cascade. While the ISCLP Kalman filter is roughly M^2 times less expensive than both reference algorithms, where M denotes the number of microphones, it is shown to perform at least similarly as compared to the former, and to outperform the latter. A MATLAB implementation is available.**

Manuscript received June 15, 2019; revised November 26, 2019; accepted January 4, 2020. Date of publication January 15, 2020; date of current version January 30, 2020. This work was supported in part by the ESAT Laboratory of KU Leuven, in the frame of KU Leuven internal fund C2-16-00449; VLAIO O&O Project no. HBC.2017.0358; EU FP7-PEOPLE Marie Curie Initial Training Network funded by the European Commission under Grant Agreement no. 316969; the European Union's Horizon 2020 research and innovation program/ERC Consolidator Grant no. 773268. This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andy W. H. Khong. *(Corresponding author: Thomas Dietzen.)*

T. Dietzen and T. van Waterschoot are with the STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics and ETC Technology Cluster Electrical Engineering, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium (e-mail: thomas.dietzen@esat.kuleuven.be; toon.vanwaterschoot@esat.kuleuven.be).

S. 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).

M. Moonen is with the ESAT, STADIUS, Katholieke Universiteit Leuven, 3001 Leuven, Belgium (e-mail: marc.moonen@esat.kuleuven.be).

Digital Object Identifier 10.1109/TASLP.2020.2966869

***Index Terms*—Dereverberation, interfering speech cancellation, noise reduction, beamforming, multi-channel linear prediction (MCLP), Kalman filter.**

I. Introduction

In many wide-spread speech processing applications such as hands-free telephony and distant automatic speech recognition, reverberation as well as additive noise and/or interfering speech impinging on a microphone may deteriorate the quality and intelligibility of the speech recordings [1]. The demanding tasks of dereverberation, noise reduction and/or interfering speech cancellation, and in particular the conjunction of these therefore remain a subject of ongoing research, with multi-microphone-based approaches exploiting spatial diversity receiving particular interest [2]–[22]. In this context, we below briefly discuss two broad concepts in multi-microphone speech enhancement, namely spatial filtering and deconvolution.

As a spatial filtering technique, beamforming is commonly used in noise reduction and interfering speech cancellation, but may as well be applied for dereverberation [2]–[4]. In order to perform both dereverberation and noise reduction, several beamforming schemes have been proposed. In [2], a cascaded approach is presented, using data-independent, superdirective beamforming for dereverberation, and data-dependent, e.g., minimum-variance distortionless response (MVDR) beamforming, for noise reduction. The generalized sidelobe canceler (GSC), a popular implementation of the MVDR beamformer, has been applied in different constellations [3], [4]. In [3], joint dereverberation and noise reduction is performed using a single GSC, while in [4], a nested structure is proposed, employing an inner GSC for dereverberation and an outer GSC for noise reduction. The GSC is composed of two parallel signal paths: a reference path and a sidelobe-cancellation (SC) path. The reference path traditionally employs a matched filter (MF), while the SC path cascades a blocking matrix (BM), blocking either the entire or the early-reverberant speech component, and an SC filter, minimizing the output power and thereby suppressing residual nuisance components in the reference path, i.e. either residual noise or both residual noise and reverberation components.

2329-9290 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

As a deconvolution technique, multi-channel linear prediction (MCLP) [5]–[22] recently prevailed in blind speech dereverberation, while noise reduction is not targeted. As opposed to beamforming, MCLP does not require spatial information on the speech source. Instead, for each microphone, the reverberation component to be canceled is modeled as a linear prediction (LP) component, i.e. as a filtered version of the delayed microphone signals, with the LP filter to be estimated. Besides iterative LP filter estimation approaches such as [6], [8], [9], [11]–[13], also adaptive approaches based on recursive least squares (RLS) [7], [10], [16], [19] as well as the Kalman filter [14], [15], [17] have been proposed in the past years. In order to reduce noise after dereverberation, multiple-output MCLP has been cascaded with MVDR beamforming in [12], [13], which was seen to be a commonly adopted approach in the 2018 CHiME-5 challenge [23]. In [22], the cascade in [12], [13] is unified. In [18], joint MCLP-based dereverberation and noise reduction is performed using a pair of alternating Kalman filters respectively estimating the LP filter and the noise-free reverberant speech component.

In [21], we have presented a comparative analysis of the GSC and MCLP. In another previous paper [20], instead of cascading MCLP and beamforming or relying on beamforming only, we have proposed to integrate the GSC and MCLP by employing an SC path and LP path in parallel, resulting in an architecture we refer to as integrated sidelobe cancellation and linear prediction (ISCLP). Within this novel architecture, we have estimated the SC and LP filters jointly by means of a single Kalman filter. Here, the spatial pre-processing blocks MF and BM require an estimate of the relative early transfer functions (RETFs), cf. also [3], while the Kalman filter requires an estimate of the power spectral density (PSD) of the desired early speech component, cf. also [14], [15], [17]. In this paper, the work in [20] is extended in the following manner. We generalize the short-time Fourier transform (STFT) domain-based signal model, which now comprises several reverberant speech components, whereof some are to be dereverberated and others to be canceled, as well as a diffuse (e.g., babble) noise component to be suppressed. This generalized acoustic scenario necessitates (non-stationary) multi-source early PSD estimation and RETF updates, which is achieved by means of the algorithm recently proposed in [24]. We further augment the proposed approach by a spectral Wiener gain post-processor, which is shown to relate to the Kalman filter's posterior state estimate. In order to demonstrate the effectiveness of the ISCLP Kalman filter, we compare against two state-of-the-art approaches – first the previously mentioned alternating Kalman filters in [18], and second a MCLP+GSC Kalman filter cascade, conceptually relating to [12], [13]. As compared to these two reference algorithms, the ISCLP Kalman filter is computationally roughly M^2 times less expensive, where M denotes the number of microphones. Yet, the ISCLP Kalman filter is shown to perform similarly as compared to the alternating Kalman filters, and to outperform the MCLP+GSC Kalman filter cascade. A MATLAB implementation and audio examples are available at [25].

The paper is organized as follows. In Section II, we present the signal model in the STFT domain. In Section III, the ISCLP Kalman filter is described. Implementational aspects are discussed in Section IV, followed by simulations in Section V.

II. Signal Model

Throughout the paper, we use the following notation: vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{I}$ and $\mathbf{0}$ denote an identity and zero matrix, $\mathbf{1}$ denotes a vector of ones, $\mathbf{A}^*$, $\mathbf{A}^T$, $\mathbf{A}^H$, and $\mathrm{E}[\mathbf{A}]$ denote the complex conjugate, the transpose, the complex conjugate transpose or Hermitian, and the expected value of a matrix $\mathbf{A}$. The operation $\mathrm{Diag}[\mathbf{a}]$ creates a diagonal matrix with the elements of $\mathbf{a}$ on its diagonal, and $\mathrm{tr}[\mathbf{A}]$ denotes the trace of $\mathbf{A}$. Submatrices are referenced either by index ranges or alternatively by sets of indices, e.g., the submatrix of $\mathbf{A}$ spanning all rows and the columns j_1 to j_2 is denoted as $[\mathbf{A}]_{:,j_1:j_2}$, and the submatrix composed of all rows and the columns of $\mathbf{A}$ with indices in the ordered set T is denoted as $[\mathbf{A}]_{:,\in T}$.

In the short-time Fourier transform (STFT) domain, with l and k indexing the frame and the frequency bin, respectively, let $y_m(l,k)$ with $m = 1,\dots,M$ denote the mth microphone signal, with M the number of microphones. In the following, we treat all frequency bins independently and hence omit the frequency index. We define the stacked microphone signal vector $\mathbf{y}(l) \in \mathbb{C}^M$,

$$\mathbf{y}(l) = (y_1(l) \ \ \cdots \ y_M(l))^T \tag{1}$$

composed of the mutually uncorrelated reverberant speech components $\mathbf{x}_n(l)$ with $n = 1,\dots,N$ originating from $N < M$ point speech sources and the noise component $\mathbf{v}(l)$, defined similarly to (1), i.e.

$$\mathbf{y}(l) = \sum_{n=1}^{N} \mathbf{x}_n(l) + \mathbf{v}(l). \tag{2}$$

Here, the reverberant speech components $\mathbf{x}_n(l)$ may be decomposed into the early and late-reverberant speech components $\mathbf{x}_{n|\mathrm{e}}(l)$ and $\mathbf{x}_{n|\ell}(l)$, i.e.

$$\mathbf{x}_n(l) = \mathbf{x}_{n|\mathrm{e}}(l) + \mathbf{x}_{n|\ell}(l), \tag{3}$$

which are commonly parted by the arrival time of the therein contained reflections and assumed to have distinct spatio-temporal properties as outlined below. Let $\mathbf{x}_{\mathrm{e}}(l) = \sum_{n=1}^{N} \mathbf{x}_{n|\mathrm{e}}(l)$ and $\mathbf{x}_{\ell}(l) = \sum_{n=1}^{N} \mathbf{x}_{n|\ell}(l)$ denote the sum of the early and late-reverberant speech components, respectively, such that $\mathbf{y}(l)$ in (2)–(3) may alternatively be written as

$$\mathbf{y}(l) = \mathbf{x}_{\mathrm{e}}(l) + \mathbf{x}_{\ell}(l) + \mathbf{v}(l). \tag{4}$$

Early reflections are assumed to arrive within the same frame,[1] where the early components in $\mathbf{x}_{n|\mathrm{e}}(l)$ are related by the RETFs in $\mathbf{h}_n(l) \in \mathbb{C}^M$ as $\mathbf{x}_{n|\mathrm{e}}(l) = \mathbf{h}_n(l)s_n(l)$. Here, without loss of generality, the RETFs are assumed to be relative to the first microphone, i.e. $[\mathbf{h}_n(l)]_1 = 1$, and $s_n(l) = [\mathbf{x}_{n|\mathrm{e}}(l)]_1$ denotes the early component in the first microphone originating from the nth source, in the following referred to as early source image. We stack $\mathbf{h}_n(l)$ and $s_n(l)$ into $\mathbf{H}(l) \in \mathbb{C}^{M\times N}$ and $\mathbf{s}(l) \in \mathbb{C}^N$,

[1] i.e. the frame length directly relates to the definition of early reflections in terms of their arrival time. In our implementation, we use a frame length of 32 ms cf. Section V-D, and hence consider reflections with a propagation delay of up to 32 ms as early.

respectively, i.e.

$$\mathbf{H}(l) = (\mathbf{h}_1(l) \quad \cdots \quad \mathbf{h}_N(l)), \tag{5}$$

$$\mathbf{s}(l) = (s_1(l) \quad \cdots \quad s_N(l))^T, \tag{6}$$

such that $\mathbf{x}_\mathrm{e}(l)$ is expressed by

$$\mathbf{x}_\mathrm{e}(l) = \mathbf{H}(l)\mathbf{s}(l). \tag{7}$$

In the following, let $N_T \leq N$ early speech source images $s_n(l)$ be defined as the target source images, and let T denote the set of the corresponding $|T| = N_T$ target-source indices. Let T' denote the complement set to T, with $|T'| = N - N_T$. In order to distinguish the target components in $\mathbf{y}(l)$ as well as their complements, we introduce the short-hand notations similar to (5)–(7),

$$\mathbf{H}_T(l) = [\mathbf{H}(l)]_{:,\in T}, \tag{8}$$

$$\mathbf{s}_T(l) = [\mathbf{s}(l)]_{\in T}, \tag{9}$$

$$\mathbf{x}_{\mathrm{e}|T}(l) = \mathbf{H}_T(l)\mathbf{s}_T(l), \tag{10}$$

and $\mathbf{H}_{T'}(l)$, $\mathbf{s}_{T'}(l)$, and $\mathbf{x}_{\mathrm{e}|T'}(l)$ similarly, such that $\mathbf{x}_\mathrm{e}(l)$ in (4) becomes

$$\mathbf{x}_\mathrm{e}(l) = \mathbf{x}_{\mathrm{e}|T}(l) + \mathbf{x}_{\mathrm{e}|T'}(l). \tag{11}$$

Our objective is to estimate

$$s_T(l) = \sum_{n\in T} s_n(l) = \mathbf{1}^T\mathbf{s}_T(l), \tag{12}$$

from $\mathbf{y}(l)$ by means of the ISCLP Kalman filter. To this end, we rely on assumptions on the spatio-temporal behavior of the individual microphone signal components. We assume that $s_n(l)$ is temporally uncorrelated across frames, i.e. we have $\mathrm{E}[s_n(l-l')s_n^*(l)] = 0$ for $l' > 0$, and with $\mathbf{x}_{n|\mathrm{e}}(l) = \mathbf{h}_n(l)s_n(l)$ consequently

$$\mathrm{E}[\mathbf{x}_{n|\mathrm{e}}(l-l')\mathbf{x}_{n|\mathrm{e}}^H(l)] = \mathbf{0} \quad \text{for} \quad l' > 0. \tag{13}$$

For speech signals, this assumption can be considered approximately justified if the STFT window length and window shift are sufficiently large. Within the limits defined by the reverberation time, we assume that the late-reverberant speech component $\mathbf{x}_{n|\ell}(l)$ is correlated to previous early source images $s_n(l-l')$ with $l' > 0$, but not to the current early source image $s_n(l)$, i.e. we have

$$\boxed{\mathrm{E}[\mathbf{x}_{n|\mathrm{e}}(l-l')\mathbf{x}_{n|\ell}^H(l)] \neq \mathbf{0} \quad \text{for} \quad l' > 0,} \tag{14}$$

$$\mathrm{E}[\mathbf{x}_{n|\mathrm{e}}(l)\mathbf{x}_{n|\ell}^H(l)] = \mathbf{0}. \tag{15}$$

Note that (14) is always satisfied practice, where the frame length is commonly much shorter than the room impulse response (RIR). Assumption (14) also implies $\mathrm{E}[\mathbf{x}_{n|\ell}(l-l')\mathbf{x}_{n|\ell}^H(l)] \neq \mathbf{0}$ for all l', i.e. we may predict $\mathbf{x}_{n|\ell}(l)$ from $\mathbf{x}_n(l-l')$, which indeed is the fundamental assumption of MCLP-based dereverberation [5]–[19]. Assumption (15) is commonly used in dereverberation [26], [27], and implicitly requires (13) to hold if (14) is assumed. Assumptions (13) and (15) allow for unbiased filter estimation [21] in MCLP-based dereverberation [5]–[19] and GSC-based dereverberation and noise reduction [3], [4], respectively. Hence, all three assumptions (13)–(15) are equally essential in the derivation of the ISCLP Kalman filter, cf. Section III. In Section IV-A, we outline how the ISCLP Kalman filter offers robustness against model deficiencies, e.g., for the case where these assumptions are violated.

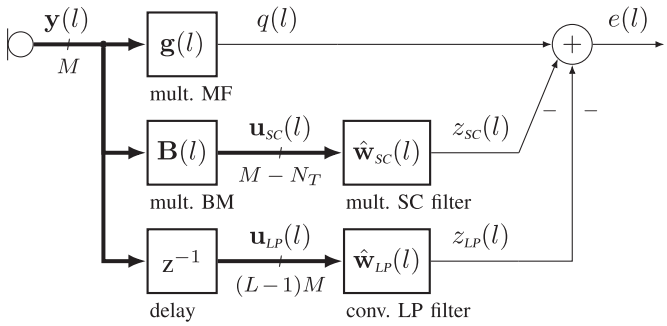


Fig. 1. The integrated sidelobe cancellation and linear prediction (ISCLP) architecture.

Similarly to $s_n(l)$, the noise component $\mathbf{v}(l)$ is assumed to be temporally uncorrelated, i.e.

$$\mathrm{E}[\mathbf{v}(l-l')\mathbf{v}^H(l)] = \mathbf{0} \quad \text{for} \quad l' > 0, \tag{16}$$

and is therefore not predictable.

Within frame l, i.e. for $l' = 0$, we further make assumptions on the spatial behavior of $\mathbf{x}_{n|\ell}(l)$ and $\mathbf{v}(l)$, namely that both may be modeled as spatially diffuse. However, as these assumptions are irrelevant in the derivation of the ISCLP Kalman filter itself, cf. Section III, but required only for parameter estimation based on [24], i.e. the estimation of the RETFs $\mathbf{H}_T(l)$ and the PSD $\varphi_{s_T}(l) = \mathrm{E}[s_T(l)s_T^*(l)]$, we treat them in the corresponding section only, cf. Section IV-B.

III. Integrated Sidelobe Cancellation and Linear Prediction Kalman Filter

We strive to estimate the target component $s_T(l)$ from the microphone signals $\mathbf{y}(l)$ defined in Section II. For this purpose, we introduce the ISCLP architecture. In Section III-A, we describe the SC and LP signal paths and filter constellations, which require spatio-temporal pre-processing of $\mathbf{y}(l)$. In Section III-B, striving for recursive filter estimation, we define an ISCLP state-space model for the SC and the LP filter, whereof a Kalman filter is deduced. The Kalman filter yields a (prior) estimate $e(l) = \hat{s}_T(l)$ of $s_T(l)$, which may further be spectrally post-processed, as shown in Section III-C.

A. ISCLP Signal Path Architecture

A block-diagram of the ISCLP architecture is depicted in Fig. 1. It integrates the GSC and MCLP and hence consists of three signal paths: a reference path employing an MF, an SC path, composed of a BM and an SC filter, and a LP path, composed of a delay and an LP filter. While the MF, the BM and the SC filter are multiplicative (mult.), i.e. they operate on a single frame, the LP filter is convolutive (conv.), i.e. it operates across frames. The MF and the BM perform spatial pre-processing, serving

unconstrained estimation of the SC filter, while the delay may analogously be considered as temporal pre-processing, serving unconstrained estimation of the LP filter. Structurally, one may interpret ISCLP either as MCLP with the conventional reference channel selection replaced by a GSC, or alternatively as a GSC employing a generalized BM (composed of a traditional BM and a delay line), and a convolutive filter (composed of the SC and the LP filter). In the following, we formally discuss the individual signal paths.

In oder to maintain the target component $s_T(l)$ in (12), the MF $\mathbf{g} \in \mathbb{C}^M$ must satisfy [28], [29]

$$\mathbf{g}^H(l)\mathbf{H}_T(l) = \mathbf{1}^T, \tag{17}$$

where a commonly used [28], [29] choice for $\mathbf{g}(l)$ adhering to (17) is

$$\mathbf{g}(l) = \mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}\mathbf{1}, \tag{18}$$

with $\mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}$ the pseudoinverse of $\mathbf{H}_T^H(l)$. In practice, we hence require an estimate $\hat{\mathbf{H}}_T(l)$ of $\mathbf{H}_T(l)$, cf. also Section IV-B. With $\mathbf{y}(l)$ as in (4), combining (10)–(12), the MF output $q(l)$ becomes

$$\begin{aligned}\boxed{q(l) = \mathbf{g}^H(l)\mathbf{y}(l)}\\ = s_T(l) + \mathbf{g}^H(l)\big(\mathbf{x}_{e|T'}(l) + \mathbf{x}_\ell(l) + \mathbf{v}(l)\big).\end{aligned} \tag{19}$$

The BM $\mathbf{B}(l) \in \mathbb{C}^{M\times M-N_T}$ must be orthogonal to $\mathbf{H}_T(l)$, i.e.

$$\mathbf{B}^H(l)\mathbf{H}_T(l) = \mathbf{0}, \tag{20}$$

and with (18) hence $\mathbf{B}^H(l)\mathbf{g}(l) = \mathbf{0}$. One may, e.g., choose $\mathbf{B}(l)$ based on the first $M - N_T$ columns of the rank-$(M - N_T)$ projection matrix to the null space of $\mathbf{H}_T(l)$ [29], i.e.

$$\mathbf{B}(l) = \big[\mathbf{I} - \mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}\mathbf{H}_T^H(l)\big]_{:,1:M-N_T}, \tag{21}$$

with $\mathbf{H}_T(l)\big(\mathbf{H}_T^H(l)\mathbf{H}_T(l)\big)^{-1}\mathbf{H}_T^H(l)$ the projection matrix to the column space of $\mathbf{H}_T(l)$. With $\mathbf{y}(l)$ as in (4), combining (10)–(11), the SC-filter input $\mathbf{u}_{SC}(l) \in \mathbb{C}^{M-N_T}$ is then given by

$$\begin{aligned}\boxed{\mathbf{u}_{SC}(l) = \mathbf{B}^H(l)\mathbf{y}(l)}\\ = \mathbf{B}^H(l)\big(\mathbf{x}_{e|T'}(l) + \mathbf{x}_\ell(l) + \mathbf{v}(l)\big),\end{aligned} \tag{22}$$

whereby the target component $\mathbf{x}_{e|T}(l) = \mathbf{H}_T(l)\mathbf{s}_T(l)$ is canceled. Using a delay of one[2] frame, the LP-filter input $\mathbf{u}_{LP}(l) \in \mathbb{C}^{(L-1)M}$ is defined by stacking $\mathbf{y}(l)$ over the past $L-1$ frames, i.e.

$$\boxed{\mathbf{u}_{LP}(l) = (\mathbf{y}^T(l-1) \quad \cdots \quad \mathbf{y}^T(l-L+1))^T.} \tag{23}$$

With the SC filter $\hat{\mathbf{w}}_{SC}(l) \in \mathbb{C}^{M-N_T}$ and the LP filter $\hat{\mathbf{w}}_{LP}(l) \in \mathbb{C}^{(L-1)M}$, the enhanced signal $e(l) = \hat{s}_T(l)$ at the output of ISCLP, also referred to as error signal in the remainder, is given by

$$e(l) = \hat{s}_T(l) = q(l) - z_{SC}(l) - z_{LP}(l), \tag{24}$$

$$\text{with} \quad z_{SC}(l) = \hat{\mathbf{w}}_{SC}^H(l)\mathbf{u}_{SC}(l), \tag{25}$$

$$z_{LP}(l) = \hat{\mathbf{w}}_{LP}^H(l)\mathbf{u}_{LP}(l). \tag{26}$$

At this point, given $q(l)$, $\mathbf{u}_{SC}(l)$, and $\mathbf{u}_{LP}(l)$, our task consists in obtaining the filters $\hat{\mathbf{w}}_{SC}(l)$ and $\hat{\mathbf{w}}_{LP}(l)$ as estimates of some yet to be defined associated true states $\mathbf{w}_{SC}(l)$ and $\mathbf{w}_{LP}(l)$, cf. Sec III-B. In this respect, let us first discuss the mutual relations between the target component $s_T(l)$ in $q(l)$ and the signals $\mathbf{u}_{SC}(l)$ and $\mathbf{u}_{LP}(l)$, as well as the consequences thereof for the filter estimation. Note that due to the delay in the LP path, the filter estimates $\hat{\mathbf{w}}_{SC}(l)$ and $\hat{\mathbf{w}}_{LP}(l)$ do not operate on the same input-data frame at the same time. The SC-filter input $\mathbf{u}_{SC}(l)$ in (22) depends on the current frame $\mathbf{y}(l)$ only, such that $\hat{\mathbf{w}}_{SC}(l)$ will exploit spatial correlations within the current frame. Due to the cancellation of $\mathbf{x}_{e|T}(l)$ at the BM output and (15), we have $\mathrm{E}[\mathbf{u}_{SC}(l)s_T^*(l)] = \mathbf{0}$. This allows for unconstrained, recursive estimation of $\mathbf{w}_{SC}(l)$, which is indeed the general incentive behind the usage of GSC-like structures [28], [29]. In contrast, the LP-filter input $\mathbf{u}_{LP}(l)$ in (23) depends on the $L-1$ previous frames $\mathbf{y}(l-l')$ with $l' = 1, \ldots, L-1$, such that $\hat{\mathbf{w}}_{LP}(l)$ will exploit spatio-temporal correlations between the current and the previous frames (but not within the current frame). Due to this delay and (13), we have $\mathrm{E}[\mathbf{u}_{LP}(l)s_T^*(l)] = \mathbf{0}$, likewise allowing for unconstrained, recursive estimation of $\mathbf{w}_{LP}(l)$. However, with both $\mathbf{u}_{SC}(l)$ and $\mathbf{u}_{LP}(l)$ containing (late-)reverberant components, the two inputs are *not* independent, i.e.

$$\boxed{\mathrm{E}[\mathbf{u}_{LP}(l)\mathbf{u}_{SC}^H(l)] \neq \mathbf{0},} \tag{27}$$

cf. (14), and as a consequence also $\mathrm{E}[z_{LP}(l)z_{SC}^*(l)] \neq 0$. In other words, a change in $\hat{\mathbf{w}}_{SC}(l)$ requires a change in $\hat{\mathbf{w}}_{LP}(l)$, and vice versa. We therefore strive to *jointly* estimate both filters.

B. ISCLP State-Space Model and Kalman Filter Update

In order to recursively estimate the SC and LP filter, we employ a Kalman filter [31]–[33], which has also been applied successfully to MCLP in previous works [14], [15], [17], [18]. Hereby, we interpret $\hat{\mathbf{w}}_{SC}(l)$ and $\hat{\mathbf{w}}_{LP}(l)$ as estimates of the true states $\mathbf{w}_{SC}(l)$ and $\mathbf{w}_{LP}(l)$, which are defined by a state-space model comprising the so-called measurement equation and the process equation. In the following, we first define the state-space model, and then present the corresponding Kalman filter update equations, which recursively estimate the true state.

As we intend to estimate $\mathbf{w}_{SC}(l)$ and $\mathbf{w}_{LP}(l)$ jointly, cf. Section III-A, we stack the SC and LP filter path into $\mathbf{u}(l) \in \mathbb{C}^{LM-N_T}$ and $\mathbf{w}(l) \in \mathbb{C}^{LM-N_T}$, i.e.

$$\mathbf{u}(l) = (\mathbf{u}_{SC}^T(l) \quad \mathbf{u}_{LP}^T(l))^T, \tag{28}$$

$$\mathbf{w}(l) = (\mathbf{w}_{SC}^T(l) \quad \mathbf{w}_{LP}^T(l))^T. \tag{29}$$

and $\hat{\mathbf{w}}(l)$ defined similarly to (29). The true state $\mathbf{w}(l)$ is considered a random variable with zero mean and correlation

[2]In MCLP literature, delays of more than one frame are commonly used [8]–[13], [15], [16], [18], [19] in order to avoid temporal target component leakage due to overlapping windows in the STFT processing, cf. Section IV-A. As we here also consider interfering reverberant speech components to be canceled, larger delays in the LP filter path however call for a convolutive SC filter [21] instead. The here proposed design did not show to be sensitive to leakage effects, cf. Section IV-A and Section V.

matrix $\mathbf{\Psi}_w(l) = \mathrm{E}[\mathbf{w}(l)\mathbf{w}^H(l)]$. We assume that $\mathbf{w}(l)$ leads to complete cancellation[3] of $\mathbf{g}^H(l)\big(\mathbf{x}_{e|T'}(l) + \mathbf{x}_\ell(l) + \mathbf{v}(l)\big)$, and therefore yielding $e(l) = s_T(l)$, cf. (19) and (24)–(26). Reformulating (24)–(26) using (28)–(29), inserting $e(l) = s_T(l)$ and rearranging yields the so-called measurement equation,

$$\boxed{q^*(l) = \mathbf{u}^H(l)\mathbf{w}(l) + s_T^*(l).} \tag{30}$$

In Kalman filter terminology, we refer to $q^*(l)$ as the measurement and to $s_T^*(l)$ as the (presumed zero-mean Gaussian.[4] and temporally uncorrelated, cf. also Section II) measurement noise with PSD $\varphi_{s_T}(l) = \mathrm{E}[s_T(l)s_T^*(l)]$. In practice, in order to implement the Kalman filter update equations, an estimate $\hat{\varphi}_{s_T}(l)$ of $\varphi_{s_T}(l)$ is required, cf. Section IV-B.

The true state $\mathbf{w}(l)$ is assumed time-varying, which accounts for potential time variations in the room impulse responses, e.g., caused by time-varying source and microphone-array positions, as well as time-varying activity of individual sources and noise powers. The so-called process equation models the evolution of the true state $\mathbf{w}(l)$ in the form of a first-order difference equation, i.e.

$$\boxed{\mathbf{w}(l) = \mathbf{A}^H(l)\mathbf{w}(l-1) + \mathbf{w}_\Delta(l).} \tag{31}$$

where $\mathbf{A}(l)$ models the state transition from one frame to the next, and the process noise $\mathbf{w}_\Delta(l)$ models a random (presumed zero-mean Gaussian and temporally uncorrelated) variation component with correlation matrix $\mathbf{\Psi}_{w_\Delta}(l) = \mathrm{E}[\mathbf{w}_\Delta(l)\mathbf{w}_\Delta^H(l)]$. Lacking deeper knowledge on the exact evolution of the true state, both $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ are commonly considered design parameters to be tuned [15], [17], [18], [34], cf. Section IV-C.

The true state $\mathbf{w}(l)$ modeled by (30)–(31) may be estimated recursively by means of the Kalman filter update equations [31], [33], which are commonly presented as two distinct sets of updates per recursion, namely an a-priori time update reflecting the state evolution, cf. (31), and an a-posteriori measurement update reflecting the current measurement, cf. (30). Specifically, let $\hat{\mathbf{w}}(l)$ and $\hat{\mathbf{w}}^+(l)$ denote the yet to be defined prior and posterior state estimates of $\mathbf{w}(l)$, respectively, and let $\tilde{\mathbf{w}}(l)$ and $\tilde{\mathbf{w}}^+(l)$ denote the associated state estimation errors, i.e.

$$\tilde{\mathbf{w}}(l) = \hat{\mathbf{w}}(l) - \mathbf{w}(l), \tag{32}$$

$$\tilde{\mathbf{w}}^+(l) = \hat{\mathbf{w}}^+(l) - \mathbf{w}(l), \tag{33}$$

with the associated state estimation error correlation matrices $\mathbf{\Psi}_{\tilde{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}^+(l)$. Then, based upon (31) and (30), respectively, the prior and posterior state estimates $\hat{\mathbf{w}}(l)$ and $\hat{\mathbf{w}}^+(l)$ shall recursively minimize the expected squared Euclidian norm of the associated state estimation error, i.e. $\mathrm{E}\big[\|\tilde{\mathbf{w}}(l)\|^2\big] = \mathrm{tr}[\mathbf{\Psi}_{\tilde{\mathbf{w}}}(l)]$ and $\mathrm{E}\big[\|\tilde{\mathbf{w}}^+(l)\|^2\big] = \mathrm{tr}[\mathbf{\Psi}_{\tilde{\mathbf{w}}}^+(l)]$. This leads to the celebrated Kalman filter update equations [31]–[33],

$$\hat{\mathbf{w}}(l) = \mathbf{A}^H(l)\hat{\mathbf{w}}^+(l-1), \tag{34}$$

$$\mathbf{\Psi}_{\tilde{w}}(l) = \mathbf{A}^H(l)\mathbf{\Psi}_{\tilde{w}}^+(l-1)\mathbf{A}(l) + \mathbf{\Psi}_{w_\Delta}(l), \tag{35}$$

$$e^*(l) = q^*(l) - \mathbf{u}^H(l)\hat{\mathbf{w}}(l), \tag{36}$$

$$\varphi_e(l) = \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l) + \varphi_{s_T}(l), \tag{37}$$

$$\mathbf{k}(l) = \mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)\varphi_e^{-1}(l), \tag{38}$$

$$\hat{\mathbf{w}}^+(l) = \hat{\mathbf{w}}(l) + \mathbf{k}(l)e^*(l), \tag{39}$$

$$\mathbf{\Psi}_{\tilde{w}}^+(l) = \mathbf{\Psi}_{\tilde{w}}(l) - \mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l), \tag{40}$$

where the time and the measurement update are given by (34)–(35) and (39)–(40), respectively. In the time update, cf. (34)–(35), the previously acquired posterior quantities $\hat{\mathbf{w}}^+(l-1)$ and $\mathbf{\Psi}_{\tilde{w}}^+(l-1)$ are propagated according to the evolution of the state $\mathbf{w}(l)$, cf. (31), yielding the prior quantities $\hat{\mathbf{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}(l)$. Then, given $\hat{\mathbf{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}(l)$, the complex conjugate error signal $e^*(l)$, its PSD $\varphi_e(l)$, and the Kalman gain $\mathbf{k}(l)$ are computed, cf. (36)–(38), thereby leveraging new information in terms of the measurement $q^*(l)$ and the measurement noise PSD $\varphi_{s_T}(l)$, cf. (30). Finally, in the measurement update, cf. (39)–(40), $e^*(l)$ and $\mathbf{k}(l)$ are utilized to update $\hat{\mathbf{w}}(l)$ and $\mathbf{\Psi}_{\tilde{w}}(l)$, yielding the posterior quantities $\hat{\mathbf{w}}^+(l)$ and $\mathbf{\Psi}_{\tilde{w}}^+(l)$. The error signal $e(l)$ in (36) thereby represents the Kalman filter estimate of $s_T(l)$, cf. also (24)–(26). As the Kalman filter minimizes $\mathrm{tr}[\mathbf{\Psi}_{\tilde{\mathbf{w}}}(l)]$ during convergence, it is easily seen that also $\varphi_e(l) = \mathrm{E}\big[|e(l)|^2\big]$ in (37) is minimized. The Kalman filter requires initialization, which we consider in Section IV-C.

C. Posterior-Like Spectral Post-Processing

With $\hat{\mathbf{w}}(l)$ a prior estimate of $\mathbf{w}(l)$, we may consider $e(l) = \hat{s}_T(l)$ in (36) a prior estimate of $s_T(l)$. After the measurement update in (39), yielding the posterior estimate $\hat{\mathbf{w}}^+(l)$ of $\mathbf{w}(l)$, we may accordingly define a posterior estimate $e^+(l) = \hat{s}_T^+(l)$ similar to (36) by

$$e^{*|+}(l) = \hat{s}_T^{*|+}(l) = q^*(l) - \mathbf{u}^H(l)\hat{\mathbf{w}}^+(l). \tag{41}$$

Interestingly, $e^+(l)$ in (41) can be shown to be a spectrally post-processed version of $e(l)$. Precisely, inserting (39) while using (36), inserting (38) and finally (37), we find

$$\begin{aligned} e^+(l) = \hat{s}_T^+(l) &= \big(1 - \mathbf{u}^H(l)\mathbf{k}(l)\big)^* e(l) \\ &= \big(1 - \mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)\varphi_e^{-1}(l)\big)^* e(l) \\ &= \frac{\varphi_{s_T}(l)}{\varphi_e(l)} e(l), \end{aligned} \tag{42}$$

where $\gamma(l) = \varphi_{s_T}(l)/\varphi_e(l)$ can be recognized as the spectral Wiener gain minimizing $\mathrm{E}\big[|s_T(l) - \gamma(l)e(l)|^2\big]$. In practice, where we rely on potentially highly non-stationary estimates $\hat{\varphi}_{s_T}(l)$, cf. Section IV-A and Section IV-B, one may prefer slowly decaying gains for perceptual reasons [35]. Therefore, instead of using (42), we propose to alternatively define $\gamma(l)$

[3]Note that complete cancellation may not necessarily be possible, e.g., if $\mathbf{v}(l) \neq \mathbf{0}$ [21], and so the true state does not necessarily exist. Nonetheless, lacking deeper knowledge on the true system, we assume that it lies in the model set.

[4]Note that the STFT coefficients of speech are said to be super-Gaussian instead of Gaussian distributed [30] The Kalman filter is the best linear state estimator also in case of non-Gaussian noises, but better non-linear estimators may indeed exist [31], [32].

and $e^+(l)$ by

$$\gamma(l) = \max\left[\frac{\varphi_{s_T}(l)}{\varphi_e(l)}, \beta\gamma(l-1)\right], \tag{43}$$

$$e^+(l) = \hat{s}_T^+(l) = \gamma(l)e(l), \tag{44}$$

with the tuning parameter $\beta \in [0,1]$ limiting the gain decay. Note that (43)–(44) reduce to (42) for $\beta = 0$, and to (36) for $\beta = 1$ and $\gamma(0) = 1$ as initial gain, since $\varphi_{s_T}(l)/\varphi_e(l) \leq 1$ due to (37).

IV. Implementational Aspects

Kalman filters perform optimally if the assumed state-space model matches the true system [31], [33]. In a practical implementation, the here presented ISCLP Kalman filter derived from the ISCLP state-space model in (30)–(31) is subject to modeling errors, requires parameter estimation, and, where deeper knowledge on the underlying system dynamics is not available, parameter tuning. These implementational aspects are discussed in the following. In Section IV-A, we qualitatively discuss the potential target component leakage due to imperfect spatio-temporal pre-processing in ISCLP and its impact on the proposed ISCLP Kalman filter. In Section IV-B, we summarize a recently proposed approach to early PSD estimation and recursive RETF updating, which we employ in conjunction with the Kalman filter. In Section IV-C, we discuss the process equation parameter tuning and Kalman filter initialization.

A. Spatio-Temporal Target Component Leakage

The previously made assumptions that $\mathrm{E}[\mathbf{u}_{SC}(l)s_T^*(l)] = \mathbf{0}$ and $\mathrm{E}[\mathbf{u}_{LP}(l)s_T^*(l)] = \mathbf{0}$, cf. Section III-A, may not be strictly satisfied in a practical implementation, which we refer to as *target component leakage*. Leakage may occur due to the following reasons. The spatial pre-processing components MF and BM rely on spatial information in terms of the RETFs $\mathbf{H}_T(l)$, cf. (18) and (21), which needs to be estimated in practice. The estimate $\hat{\mathbf{H}}_T(l)$ commonly contains estimation errors, i.e. we have $\hat{\mathbf{H}}_T(l) \neq \mathbf{H}_T(l)$. Further, the RETF-based data model in (7) itself may be erroneous, e.g., due to dependencies across frequency bins [36]. Finally, the assumption in (15) that $\mathbf{x}_{n|\mathrm{e}}(l)$ and $\mathbf{x}_{n|\ell}^H(l)$ are uncorrelated may be violated, e.g., due to overlapping windows in the STFT processing. In general, these estimation and modeling errors cause incomplete blocking and therefore target component leakage through the BM, such that $\mathrm{E}[\mathbf{u}_{SC}(l)s_T^*(l)] \neq \mathbf{0}$, cf. (19), (22). This may be referred to as *spatial target component leakage*. Similarly, if $s_T(l)$ is temporally correlated such that (13) is violated, e.g., due to overlapping windows in the STFT processing or to too small window lengths and shifts, we find $\mathrm{E}[\mathbf{u}_{LP}(l)s_T^*(l)] \neq \mathbf{0}$, cf. (19), (23), which may be referred to as *temporal target component leakage*.

Potentially, spatial and temporal leakage cause a biased [21] filter estimate $\hat{\mathbf{w}}(l)$, which leads to partial suppression of $s_T(l)$, also referred to as speech cancellation in GSC terminology [28], [29], or excessive whitening in MCLP terminology [5]. However, note that the Kalman filter offers inherent robustness towards target-component leakage. To see this, consider the measurement update terms in (39)–(40), respectively given by $\mathbf{k}(l)e^*(l)$ and $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$. Using (30) and (32), we may express $e^*(l)$ in (36) in terms of $s_T^*(l)$, while using (37), we may similarly express $\mathbf{k}(l)$ in (38) in terms of $\varphi_{s_T}(l)$, i.e.

$$e^*(l) = s_T^*(l) - \mathbf{u}^H(l)\tilde{\mathbf{w}}(l), \tag{45}$$

$$\mathbf{k}(l) = \frac{\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l)}{\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)\mathbf{u}(l) + \varphi_{s_T}(l)}. \tag{46}$$

From (45)–(46), we note that $\varphi_{s_T}(l) = \mathrm{E}[s_T(l)s_T^*(l)]$ acts as a regularization parameter in both update terms $\mathbf{k}(l)e^*(l)$ and $\mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_{\tilde{w}}(l)$. Consequently, strong target powers inhibit the measurement update, while weak target powers promote it. Put differently, in terms of robustness towards target-component leakage and convergence, the Kalman filter benefits from non-stationarities and sparsity in $\varphi_{s_T}(l)$ across time. Note that in recursive MCLP implementations based on the weighted prediction error (WPE) criterion and RLS [7], [10], [16], [19], the target-component PSD similarly appears as a regularization term in the update equations.

In practice, we rely on estimates $\hat{\varphi}_{s_T}(l)$, which should hence maintain non-stationarities. In WPE RLS literature, the target-component PSD estimate is obtained, e.g., directly from the plain microphone signals [7], based on a late-reverberant PSD estimate obtained by means of an exponential decay model [10], [16], or using a neural network [19]. Here, as we consider a more generic signal model comprising several reverberant speech components and diffuse noise, cf. Section II, we instead estimate $\hat{\varphi}_{s_T}(l)$ by means of [24], cf. Section IV-B.

B. Target PSD Estimation and RETF Update

We require an RETF estimate $\hat{\mathbf{H}}_T(l)$ of $\mathbf{H}_T(l)$, cf. (18) and (21), and a PSD estimate $\hat{\varphi}_{s_T}(l)$ of $\varphi_{s_T}(l)$, cf. (37) and (43). To this end, we use an algorithm recently proposed in [24] by the authors of this paper, which computes early PSD estimates and recursively updates the RETF estimates for all N point sources. The algorithm [24] is summarized as follows.

Let $\mathbf{\Psi}_{x_\mathrm{e}}(l) = \mathrm{E}[\mathbf{x}_\mathrm{e}(l)\mathbf{x}_\mathrm{e}^H(l)]$ denote the correlation matrix of $\mathbf{x}_\mathrm{e}(l)$ within frame l, which generally has rank N and is given by

$$\mathbf{\Psi}_{x_\mathrm{e}}(l) = \mathbf{H}(l)\,\mathrm{Diag}[\boldsymbol{\varphi}_s(l)]\mathbf{H}^H(l), \tag{47}$$

$$\boldsymbol{\varphi}_s(l) = (\varphi_{s_1}(l) \quad \cdots \quad \varphi_{s_N}(l))^T, \tag{48}$$

with $\varphi_{s_n}(l)$ denoting the PSD of the early speech source image $s_n(l)$. Instead of directly using the conventional early correlation matrix model in (47), the algorithm in [24] is based on its factorization, i.e. it relies on the square-root model

$$\mathbf{\Psi}_{x_\mathrm{e}}^{1/2}(l)\mathbf{\Omega}(l) = \mathbf{H}(l)\,\mathrm{Diag}[\boldsymbol{\varphi}^{1/2}(l)], \tag{49}$$

where $\mathbf{\Psi}_{x_\mathrm{e}}^{1/2}(l) \in \mathbb{C}^{M\times N}$ and $\boldsymbol{\varphi}^{1/2} \in \mathbb{C}^N$ are some square roots of $\mathbf{\Psi}_{x_\mathrm{e}}(l)$ and $\boldsymbol{\varphi}_s(l)$ such that $\mathbf{\Psi}_{x_\mathrm{e}}^{1/2}(l)\mathbf{\Psi}_{x_\mathrm{e}}^{H/2}(l) = \mathbf{\Psi}_{x_\mathrm{e}}(l)$ and $\mathrm{Diag}[\boldsymbol{\varphi}^{H/2}(l)]\boldsymbol{\varphi}^{1/2}(l) = \boldsymbol{\varphi}_s(l)$, respectively, and $\mathbf{\Omega}(l)$ is a unitary matrix, i.e. $\mathbf{\Omega}(l)\mathbf{\Omega}^H(l) = \mathbf{I}$, which accounts for the non-uniqueness of both square-roots. Note that right-multiplying each side of (49) with its Hermitian yields (47). In the estimation, we distinguish the prior and posterior RETF estimates $\hat{\mathbf{H}}(l)$ and $\hat{\mathbf{H}}^+(l)$, respectively, and assume that initial RETF

estimates $\hat{\mathbf{H}}(0)$ are available, which may be based on, e.g., initial single-source RETF estimates acquired from segments with distinctly active sources [37], or some initial knowledge or estimates of the associated dicrections of arrival (DoAs) [38], [39]. Given a (to be obtained) square-root estimate $\hat{\mathbf{\Psi}}_{x_e}^{1/2}(l)$ and a prior RETF estimate $\hat{\mathbf{H}}(l)$, which is propagated from the previous posterior, i.e. $\hat{\mathbf{H}}(l) = \hat{\mathbf{H}}^{+}(l-1)$, we first obtain the unitary and diagonal estimates $\hat{\mathbf{\Omega}}(l)$ and $\mathrm{Diag}[\hat{\boldsymbol{\varphi}}^{1/2}]$, yielding $\hat{\boldsymbol{\varphi}}_s(l) = \mathrm{Diag}[\hat{\boldsymbol{\varphi}}^{H/2}]\hat{\boldsymbol{\varphi}}^{1/2}$, and based on these estimates second update the RETF estimate, yielding the posterior $\hat{\mathbf{H}}^{+}(l)$, whereat the recursion is closed. Here, both steps are based on approximation error minimization with respect to the square-root model in (49). Given $\hat{\boldsymbol{\varphi}}_s(l)$ and $\hat{\mathbf{H}}^{+}(l)$, we extract $\hat{\varphi}_{s_T}(l)$ and $\hat{\mathbf{H}}_T(l)$ as $\hat{\varphi}_{s_T}(l) = \mathbf{1}^T[\hat{\boldsymbol{\varphi}}_s(l)]_{\in T}$ and $\hat{\mathbf{H}}_T(l) = [\hat{\mathbf{H}}^{+}(l)]_{\in T}$, cf. Section II.

The said required square root $\mathbf{\Psi}_{x_e}^{1/2}(l)$ is estimated in the following manner. While $\mathbf{x}_{n|\ell}(l)$ and $\mathbf{v}(l)$ exhibit a fundamentally different temporal behavior across frames, cf. Section II, we assume that their spatial behavior within frame l is the same. Specifically, we model both $\mathbf{x}_{n|\ell}(l)$ and $\mathbf{v}(l)$ as spatially diffuse with coherence matrix $\mathbf{\Gamma} \in \mathbb{C}^{M\times M}$, which may be computed from the microphone array geometry [40], [41] and is therefore assumed to be known. For the late reverberant component $\mathbf{x}_{n|\ell}(l)$, this is a commonly made assumption [26], [27], [40]. For the noise component $\mathbf{v}(l)$, the assumption is commonly made for noise types such as, e.g., babble noise [42], which we use in our simulations, cf. Section V. Based on these assumptions, the microphone signal correlation matrix $\mathbf{\Psi}_y(l) = \mathrm{E}[\mathbf{y}(l)\mathbf{y}^H(l)]$ may be written as

$$\mathbf{\Psi}_y(l) = \mathbf{\Psi}_{x_e}(l) + \varphi_d(l)\mathbf{\Gamma}, \tag{50}$$

with $\varphi_d(l) = \sum_{n=1}^{N} \varphi_{x_{n|\ell}}(l) + \varphi_v(l)$ and $\varphi_{x_{n|\ell}}(l)$ and $\varphi_v(l)$ denoting the PSDs of the late-reverberant speech components and the diffuse noise component, respectively. We obtain a subspace representation of (50) by means of the generalized eigenvalue decomposition (GEVD) of $\mathbf{\Psi}_y(l)$ and $\mathbf{\Gamma}$. Based on the generalized eigenvectors and generalized eigenvalues, $\mathbf{\Psi}_y(l)$ may be decomposed into a diffuse component, cf. also the diffuse PSD estimator in [27], and a factorized early rank-N component $\mathbf{\Psi}_{x_e}(l) = \mathbf{\Psi}_{x_e}^{1/2}(l)\mathbf{\Psi}_{x_e}^{H/2}(l)$. A temporally smooth estimate $\hat{\mathbf{\Psi}}_{y|sm}(l)$ of $\mathbf{\Psi}_y(l)$ itself is obtained from the microphone signals by recursively averaging $\mathbf{y}^H(l)\mathbf{y}(l)$. In order to restore non-stationarities, we desmooth[5] the generalized eigenvalues of $\hat{\mathbf{\Psi}}_{y|sm}(l)$ and $\mathbf{\Gamma}$ and thereby yield non-stationary PSD estimates in the subsequent processing steps, as as favored in the Kalman filter, cf. Section IV-A. For further details, we refer the interested reader to [24].

C. Process Equation Parameter Tuning and Initialization

The tracking and convergence behavior of the Kalman filter depends on its process equation parameter tuning and initialization. The process equation models the evolution of the state by means of the parameters $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$, cf. (31) and (34)–(35). In practice, only limited knowledge of the state evolution is available, such that $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ are commonly left to tuning [15], [17], [18], [34]. Typically, both $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ are chosen to be scaled identities, with $\mathbf{A}(l)$ commonly time-invariant [15], [17], [18], [34] and acting as a forgetting factor [17], [34], and $\mathbf{\Psi}_{w_\Delta}(l)$ either time-variant [15], [18], [34] or time-invariant [17]. Here, we set $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ based on the assumption that the state correlation matrix $\mathbf{\Psi}_w(l)$ is time-invariant, i.e. $\mathbf{\Psi}_w(l) = \mathbf{\Psi}_w$. Unfortunately, $\mathbf{\Psi}_w$ is unknown and not available in practice, however, we may define a rough guess $\bar{\mathbf{\Psi}}_w$. Given such a guess $\bar{\mathbf{\Psi}}_w$, by means of a forgetting factor $\alpha \in (0,1)$, we may account for a steadily time-varying acoustic scenario and true state $\mathbf{w}(l)$ by setting

$$\mathbf{A}(l) = \sqrt{\alpha}\,\mathbf{I}, \tag{51}$$

$$\mathbf{\Psi}_{w_\Delta}(l) = (1-\alpha)\bar{\mathbf{\Psi}}_w, \tag{52}$$

such that if $\bar{\mathbf{\Psi}}_w = \mathbf{\Psi}_w$, we rightly have $\mathbf{\Psi}_w = \alpha\mathbf{\Psi}_w + (1-\alpha)\mathbf{\Psi}_w$ from (31). While $\bar{\mathbf{\Psi}}_w$ may rather be defined by design than by truly estimating $\mathbf{\Psi}_w$, the notion of $\bar{\mathbf{\Psi}}_w$ being a rough guess of $\mathbf{\Psi}_w$ may nonetheless guide its definition to some extent. Here, we choose a diagonal matrix with distinct diagonal elements. With $\bar{\mathbf{\Psi}}_w = \mathrm{Diag}[\bar{\boldsymbol{\psi}}_w]$, let $\bar{\boldsymbol{\psi}}_{w_{SC}} \in \mathbb{R}^{M-N_T}$ and $\bar{\boldsymbol{\psi}}_{w_{LP}} \in \mathbb{R}^{(L-1)M}$ denote the subvectors of $\bar{\boldsymbol{\psi}}_w$ associated to the SC and the LP filter, respectively, which we treat separately. Expecting lower values for later prediction coefficients in the LP filter, we choose the power of the diagonal elements in $\bar{\boldsymbol{\psi}}_{w_{LP}}$ to drop exponentially each M elements, i.e. we set

$$\bar{\boldsymbol{\psi}}_{w_{SC}} = \bar{\psi}_{w_{SC}}\mathbf{1}, \tag{53}$$

$$\bar{\boldsymbol{\psi}}_{w_{LP}} = \begin{pmatrix} \bar{\psi}_{w_{LP}}^{1}\mathbf{1}^T & \dots & \bar{\psi}_{w_{LP}}^{L-1}\mathbf{1}^T \end{pmatrix}^T, \tag{54}$$

with $\bar{\psi}_{w_{SC}} > 0$ and $\bar{\psi}_{w_{LP}} \in (0,1)$ further adjustable.

The matrix $\bar{\mathbf{\Psi}}_w$ may also be used to initialize the Kalman filter. With the commonly chosen initial state estimate $\hat{\mathbf{w}}(0) = \mathbf{0}$, we have $\tilde{\mathbf{w}}(0) = \mathbf{w}(0)$ in (32), such that the true initial state estimation error correlation matrix $\mathbf{\Psi}_{\tilde{w}}(0)$ becomes $\mathbf{\Psi}_{\tilde{w}}(0) = \mathbf{\Psi}_w(0) = \mathbf{\Psi}_w$. Therefore, we initialize the Kalman filter by

$$\hat{\mathbf{w}}(0) = \mathbf{0}, \tag{55}$$

$$\hat{\mathbf{\Psi}}_{\tilde{w}}(0) = \bar{\mathbf{\Psi}}_w, \tag{56}$$

in (34)–(35), where $\hat{\mathbf{\Psi}}_{\tilde{w}}(0)$ in (56) is an estimate if $\bar{\mathbf{\Psi}}_w \neq \mathbf{\Psi}_w$. Finally, note that the process equation parameter tuning in (51)–(52) may also be considered from a (re-)initialization perspective. In case of meaningful measurement updates, the Kalman filter tracks $\mathbf{w}(l)$, but otherwise tends to return to its initial condition due to (51)–(52), such that explicit re-initialization as, e.g., in case of a sudden change in the acoustic environment, is not necessary. To see this, consider the case where, e.g., $\mathbf{u}(l) = \mathbf{0}$ for a period of time, such that no measurement update is performed. In this case, regardless of their current values, we have $\hat{\mathbf{w}}(l)$ slowly converging to $\mathbf{0}$ and $\hat{\mathbf{\Psi}}_{\tilde{w}}(l)$ slowly converging to $\bar{\mathbf{\Psi}}_w$, cf. (34)–(35). Note that if desired, explicit re-initialization may still easily be incorporated in the proposed concept, namely by defining α time-variant and setting it to zero at the determined re-initialization point.

[5] Considering recursive averaging as an invertible recursive filtering operation, the generalized eigenvalues may be desmoothed by means of the corresponding inverse filter.

V. Simulations

In order to demonstrate the effectiveness of the presented ISCLP Kalman filter, we define two case studies, case A and case B. In case A, we compare to the (computationally more demanding) alternating Kalman filters proposed in [18]. Here, we consider one reverberant speech and a babble noise component, $\mathbf{x}_1(l)$ and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ containing the target component $\mathbf{x}_{\mathrm{e}|T}(l) = \mathbf{x}_{1|\mathrm{e}}(l)$. In case B, we compare to a (computationally more demanding) MCLP+GSC Kalman filter cascade, which conceptually relates to [12], [13] in that it cascades linear prediction and beamforming. Here, we consider two reverberant speech components and a babble noise component, $\mathbf{x}_1(l)$, $\mathbf{x}_2(l)$, and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ again containing the target component $\mathbf{x}_{\mathrm{e}|T}(l) = \mathbf{x}_{1|\mathrm{e}}(l)$, and $\mathbf{x}_2(l)$ an interfering speech component to be canceled. In both cases, we investigate the algorithms' behavior depending on the signal-to-noise ratio, SNR, which is defined as the power ratio of $\mathbf{x}_1(l)$ to $\mathbf{v}(l)$, and depending on the filter length L. In case A, we additionally investigate the convergence behavior.

In what follows, we describe the two reference algorithms in more detail in Section V-A, the performance measures in Section V-B, the acoustic scenario in Section V-C, the algorithmic settings in V-D, and finally the simulation results in Section V-E.

A. Reference Algorithms

We discuss the alternating Kalman filters in Section V-A1 and the MCLP+GSC Kalman filter cascade in Section V-A2.

1) Case A: Alternating Kalman Filters: In [18], MCLP-based dereverberation and noise reduction is performed in each microphone channel using two alternating Kalman filters. The Kalman filter dedicated to dereverberation estimates a multiple-output LP filter, and the Kalman filter dedicated to noise reduction estimates the noise-free reverberant speech component. The enhanced signal is computed from the posterior state estimates of both Kalman filters. The two state vectors have dimensions $M^2(L-1)$ and $M(L-1)$, respectively, while the ISCLP Kalman filter requires a single state vector with dimension $ML - N_T$ only with $N_T = 1$ in case A, cf. (29). Since the Kalman filter in general exhibits a quadratic computational cost in the state vector dimension, the alternating Kalman filters are computationally roughly M^2 times as demanding as the ISCLP Kalman filter. The two state space models do not provide a spatial distinction between point sources (and therefore do not require RETF estimates, as opposed to the ISCLP Kalman filter) and further do not consider temporally correlated interference components such as interfering reverberant speech. We hence set $\mathbf{x}_2(l) = \mathbf{0}$ when comparing to [18], i.e. interfering speech is absent, cf. Section V-C2.

The alternating Kalman filters require correlation matrix estimates of the measurement and process noises, more precisely of the random variation of the multiple-output LP filter state, comparable to $\mathbf{\Psi}_{w_\Delta}(l)$ in the ISCLP Kalman filter, cf. (31), the early component $\mathbf{\Psi}_{x_\mathrm{e}|T}(l) = \mathbf{\Psi}_{x_\mathrm{e}}(l)$, the early-plus-noise component $\mathbf{\Psi}_{x_\mathrm{e}}(l) + \mathbf{\Psi}_v(l)$, and the noise component $\mathbf{\Psi}_v(l)$ [18]. In the original implementation in [18], a time-invariant estimate $\hat{\mathbf{\Psi}}_v$ is assumed to be available, which we here compute in an oracle fashion from $\mathbf{v}(l)$ directly, while the other correlation matrices are estimated based on the previous state estimates and error signals of the alternating Kalman filters. For the sake of a fair and more meaningful comparison, we implement two versions of [18]. The first version is implemented as proposed in [18] and discussed above, subsequently referred to as the original alternating Kalman filters. In the second version, we align the parameter estimation and tuning towards the proposed approach, i.e. $\mathbf{\Psi}_{x_\mathrm{e}}(l)$ is instead estimated based on [24], cf. Section IV-B, and the process equation parameters modeling the evolution of the multiple-output LP filter state are defined similarly to Section IV-C, subsequently referred to as the modified alternating Kalman filters.

2) Case B: MCLP+GSC Kalman Filter Cascade: In [12], [13], multiple-output MCLP based on the (iterative) WPE criterion [8], [9] is cascaded with MVDR beamforming in order to reduce noise after dereverberation, which became a popular approach in the CHiME-5 challenge [23]. For the sake of a close comparison, however, we here instead compare to a (recursive) multiple-output MCLP-based Kalman filter cascaded with a (recursive) GSC-based Kalman filter, subsequently referred to as MCLP+GSC. Herein, we estimate the LP and SC filters independently. The enhanced signal at the GSC output is computed using spectral post-processing of the same kind as in (43)–(44). The two state vectors have dimensions $M^2(L-1)$ and $M-1$, respectively, while the ISCLP Kalman filter requires a single state vector with dimension $ML - N_T$ only with $N_T = 1$ in case B, cf. (29). Since the Kalman filter in general exhibits a quadratic computational cost in the state vector dimension, the MCLP+GSC Kalman filter cascade is computationally roughly M^2 times as demanding as the ISCLP Kalman filter. The GSC state space model does provide a spatial distinction between point sources (based on an RETF estimate, as the ISCLP Kalman filter). We hence set $\mathbf{x}_2(l) \neq \mathbf{0}$ when comparing to the MCLP+GSC Kalman filter cascade, i.e. interfering speech is present, cf. Section V-C2.

The MCLP and GSC Kalman filters require correlation matrix estimates of their respective measurement and process noises, more precisely of the random variation of the multiple-output LP filter and SC filter state, respectively, defined similarly to the corresponding SC and LP submatrices of $\mathbf{\Psi}_{w_\Delta}(l)$ in the ISCLP Kalman filter, cf. (31), and the early components $\mathbf{\Psi}_{x_\mathrm{e}|T}(l)$ and $\varphi_{s_T}(l)$, respectively, computed based on [24] as in the proposed ISCLP Kalman filter, cf. Section IV-B.

B. Performance Measures

As performance measures, we choose the perceptual evaluation of speech quality [43], $PESQ$, with mean opinion scores of objective listening quality $\in [1, 4.5]$, the short-time objective intelligibility [44], $STOI$, with scores $\in [0, 1]$, the frequency-weighted segmental signal-to-interference ratio [35], [45], SIR^{fws}, in dB, and the cepstral distance [35], [45], CD, in dB. While high values are preferable for $PESQ$, $STOI$, and SIR^{fws}, low values are preferred for CD. These intrusive measures require a clean reference signal $\tilde{s}_T(l)$, which approximates the target signal $s_T(l)$ in (12). In order to generate $\tilde{s}_T(l)$, we convolve the target speech source signal with

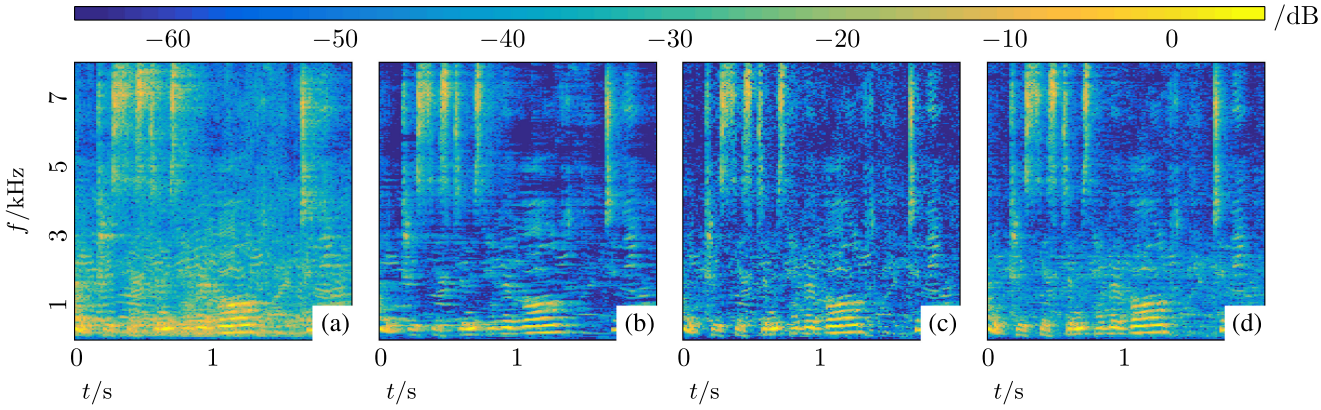


Fig. 2. Exemplary spectrograms depicting 2 s of (a) the reference microphone signal, and the corresponding outputs of (b) the original alternating Kalman filters, (c) the modified alternating Kalman filters, and (d) the ISCLP Kalman filter for $L = 6$ at $SNR = 10$ dB.

the early part of the RIR to the first microphone, cf. Section V-C, whereat we define the first N_{STFT} samples of the RIR as its early part, with N_{STFT} the analysis and synthesis window length of the STFT processing corresponding to 32 ms, cf. Section V-D. Note that due to modeling errors in the RETF-model in (7), we generally have $\tilde{s}_T(l) \neq s_T(l)$. When investigating the dependency on SNR or L, we compute the measures from 4 s to 10 s, i.e. roughly after convergence. When investigating the convergence behavior, we compute the measures within sliding windows of 2 s each. The computed measures are averaged over several individual simulations, cf. Section V-C.

C. Acoustic Scenario

We describe the acoustic scenarios without and with interfering speaker in Section V-C1 and Section V-C2, respectively.

1) Case A. Without Interfering Speech: In case A, the microphone signals are composed of one reverberant speech and a babble noise component, $\mathbf{x}_1(l)$ and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$. To generate $\mathbf{x}_1(l)$, we use RIRs to a linear microphone array, measured [46] in a room of 0.61 s reverberation time of and 0.67 m critical distance for omnidirectional sources. The linear microphone array contains $M = 5$ microphones with 8 cm inter-microphone distance. The source is positioned in 2 m distance (i.e., at roughly three times the critical distance) of the microphone array. When investigating the dependency on SNR or L, the speech source remains positioned at 0° relative to the broad-side direction during 10 s of simulation. When investigating the convergence behavior, the speech source remains positioned at 0° for the first 8 s, then jumping to 15°, where it remains for another 10 s. Both female and male speech [47] are used as speech source signals. The babble noise component is generated using [42], [48]. From the speech source signal files and the babble noise file [48], we randomly select individual segments, yielding individual simulations to be averaged in the performance evaluation, cf. Section V-B. In total, when investigating the dependency on SNR or L, we generate 64 individual simulations per condition. When investigating the convergence behavior, we generate 128 individual simulations.

2) Case B. With Interfering Speech: In case B, the microphone signals are composed of two reverberant speech components and a noise component, $\mathbf{x}_1(l)$, $\mathbf{x}_2(l)$, and $\mathbf{v}(l)$, with $\mathbf{x}_1(l)$ again containing the target component $\mathbf{x}_{e|T}(l) = \mathbf{x}_{1|e}(l)$, and $\mathbf{x}_2(l)$ an interfering speech component. We investigate the dependency on SNR and L, and generate $\mathbf{x}_1(l)$ and $\mathbf{v}(l)$ in the same manner as in case A, cf. Section V-C1. To generate $\mathbf{x}_2(l)$, we use the same set of RIR measurements, where the associated source is positioned in 2 m distance at either $\{30, 60, 90\}^\circ$. If $\mathbf{x}_1(l)$ contains female speech, then $\mathbf{x}_2(l)$ contains male speech [47] and vice versa. On average, $\mathbf{x}_1(l)$ and $\mathbf{x}_2(l)$ have roughly the same power. From the speech source signal files and the babble noise file, we randomly select individual segments, generating $3 \cdot 64 = 192$ individual simulations per condition to be averaged in the performance evaluation, cf. Section V-B.

D. Algorithmic Settings

In our simulations, the sampling frequency is $f_s = 16$ kHz, and the STFT analysis and synthesis uses square-root Hann windows of $N_{STFT} = 512$ samples with 50% overlap. When investigating the dependency on SNR and the convergence behavior, we set $L = 6$ in (23). The estimates $\hat{\boldsymbol{\varphi}}_s(l)$ and $\hat{\mathbf{H}}^+(l)$, required in (37) and (18), (21) are obtained by means of [24], cf. Section IV-B. In (51)–(52), we set α such that $10\log_{10}(1-\alpha) = -25$ dB. Expecting lower values for SC filter coefficients at higher frequencies due to generally reduced spatial correlations between individual microphones, we choose $\bar{\psi}_{w_{SC}}$ in (53) to be frequency-dependent with $10\log_{10}\bar{\psi}_{w_{SC}}$ decreasing linearly from 0 dB at 0 kHz to -15 dB at 8 kHz. In (54), we set $10\log_{10}\bar{\psi}_{w_{LP}} = -4$ dB. In (43), we set β such that $20\log_{10}\beta = -2$ dB, and $\gamma(0) = 1$.

E. Results

We discuss the results in case A and B in Section V-E1 and Section V-E2, respectively. Audio examples are available at [25].

1) Case A: Consider the spectrograms in Fig. 2 depicting 2 s of (a) the reference microphone signal $y_1(l)$, and the corresponding outputs of (b) the original alternating Kalman filters, (c) the modified alternating Kalman filters, and (d) the ISCLP Kalman filter for $L = 6$ in an exemplary simulation at

$SNR = 10$ dB. As can be seen by comparison with (a), all three algorithms in (b)–(d) considerably reduce reverberation and noise. Yet, their spectrograms exhibit slightly different features. As opposed to the modified alternating Kalman filters and the ISCLP Kalman filter (c)–(d), the original alternating Kalman filters (b) show some amount of temporal smearing resembling musical noise [18]. This is due to errors in the correlation matrix estimates used to update the alternating Kalman filters, which in turn are computed recursively based on the alternating Kalman filters' previous state estimates and error signals [18]. In contrast, in the modified alternating Kalman filters and the ISCLP Kalman filter, the required correlation matrix and PSD estimates are computed directly from the microphone signals while maintaining non-stationarities, cf. Section IV-B and Section V-A1. As compared to the modified alternating Kalman filters (c), the signal power in the ISCLP Kalman filter (d) decays somewhat less quickly after transient speech components, which is due to $\beta > 0$ in (43), cf. Section V-D, resulting in a perceptually somewhat more pleasant sound image [25].

Fig. 3 shows the performance in terms of (a) $PESQ$, (b) $STOI$, (c) SIR^{fws}, and (d) CD versus SNR for the reference microphone signal [·····], the original alternating Kalman filters [- - -], the modified alternating Kalman filters [—■—], and the ISCLP Kalman filter [—●—] with $L = 6$. In this and the following figures, the graphs denote medians over all individual simulations, cf. Section V-C, and the shaded areas indicate the range from the first to the third quartile. Overall, the measures show a high degree of agreement. As expected, the reference microphone signal reaches better scores at higher SNR values in all measures. Above roughly $SNR = -5$ dB, all three algorithms show a significant improvement over the reference microphone signal in all measures, least pronounced in $STOI$. The modified alternating Kalman filters generally outperform the original alternating Kalman filters, validating the modified parameter estimation and tuning aligned to the proposed ISCLP Kalman filter, cf. Section V-A1. In terms of $PESQ$, $STOI$, and CD, the ISCLP Kalman filter reaches very similar scores as compared to the modified alternating Kalman filters. In terms of SIR^{fws}, the ISCLP Kalman filter performs somewhat worse than the modified alternating Kalman filters above $SNR = 20$ dB, which is due to a small amount of speech cancellation caused by the SC filter, cf. Section IV-A. Note that in this SNR range, the babble noise component $\mathbf{v}(l)$ becomes negligible, i.e. reverberant interference is pre-dominant, which can be handled by the LP filter only. The SC filter therefore becomes superfluous in this case. Further simulations showed that the ISCLP Kalman filter may reach similar SIR^{fws} scores as compared to the modified alternating Kalman filters if the SC filter variance $\bar{\psi}_{w_{SC}}$ in (53) is set depdending on the SNR, which allows to essentially switch off the SC filter at high SNR values, and thereby avoid unnecessary speech cancellation.

Fig. 4 depicts the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [—■—], and the ISCLP Kalman filter [—●—] at $SNR = 25$ dB. Note that in Fig. 4 and in the following figures presenting performance improvements, the resolution of the vertical axes is twice as large as in Fig. 3. Again, the measures show a high degree of agreement. We find that in all measures, the original alternating Kalman filters generally yield less improvement and in addition show a stronger dependency on L as compared to the modified alternating Kalman filters and the ISCLP Kalman filter. The improvement for both the modified alternating Kalman filters and the ISCLP Kalman filter saturates at roughly $L = 6$. The original alternating Kalman filters reach the largest improvement between $L = 8$ and $L = 10$. In terms of (c) ΔSIR^{fws} and (d) ΔCD, however, as opposed to the other two algorithms, its performance decays again for larger values of L [18]. Further simulations showed that for all three algorithms, the dependency on L decreases with decreasing SNR values. This is expected since at low SNR values, the babble noise component $\mathbf{v}(l)$ becomes pre-dominant, which is temporally uncorrelated, cf. Section II, and may therefore not be suppressed by the LP filter.

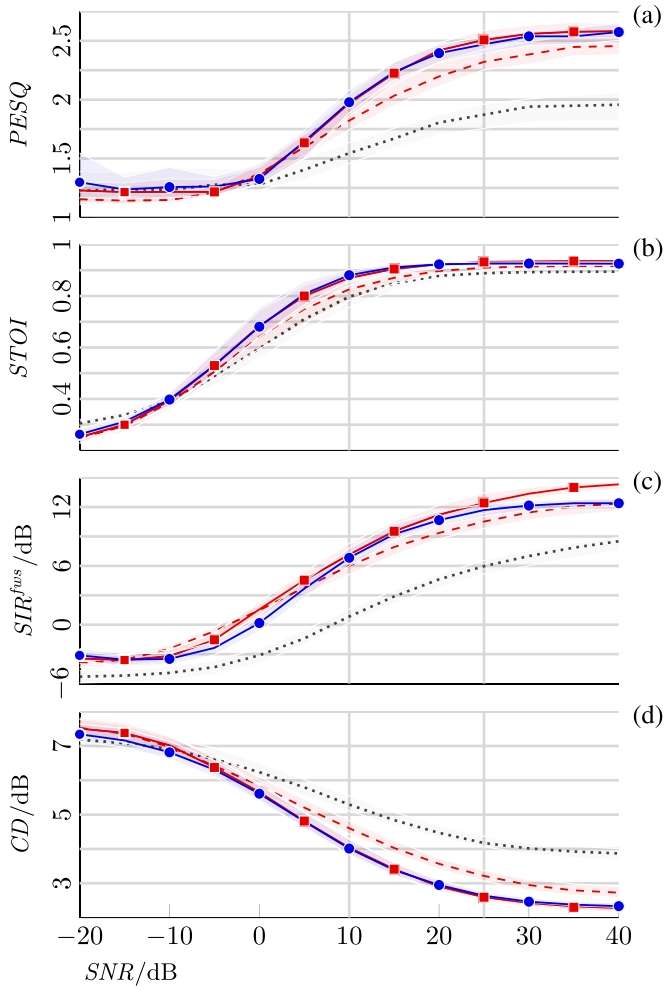


Fig. 3. (a) $PESQ$, (b) $STOI$, (c) SIR^{fws}, and (d) CD versus SNR for the reference microphone signal [·····], the original alternating Kalman filters [- - -], the modified alternating Kalman filters [—■—], and the ISCLP Kalman filter [—●—] with $L = 6$ if interfering speech is absent.

Fig. 5 shows the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus time t with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating

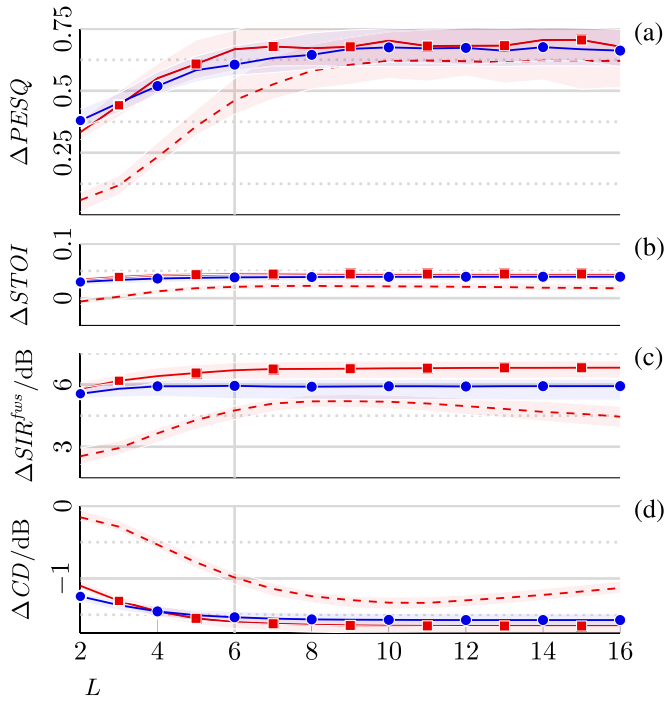


Fig. 4. (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [—■—], and the ISCLP Kalman filter [—●—] at $SNR = 25$ dB if interfering speech is absent.

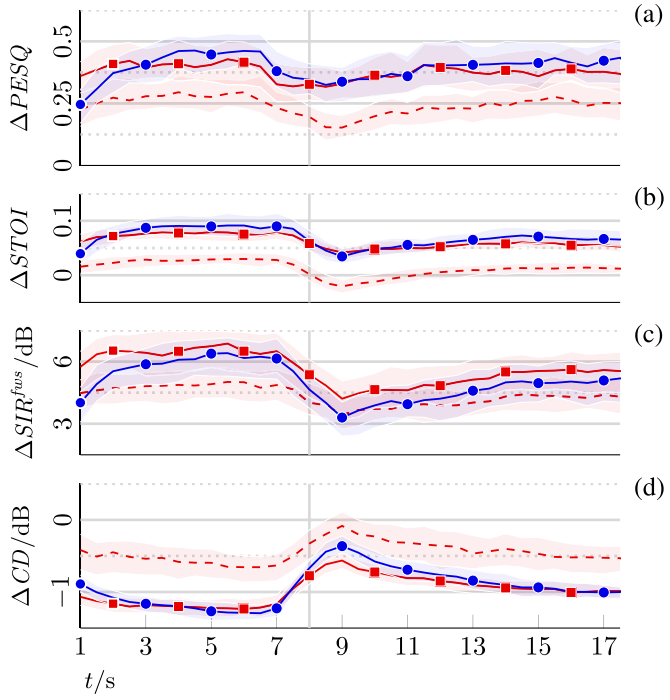


Fig. 5. (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus t with respect to the reference microphone signal for the original alternating Kalman filters [- - -], the modified alternating Kalman filters [—■—], and the ISCLP Kalman filter [—●—] with $L = 6$ at $SNR = 10$ dB if interfering speech is absent.

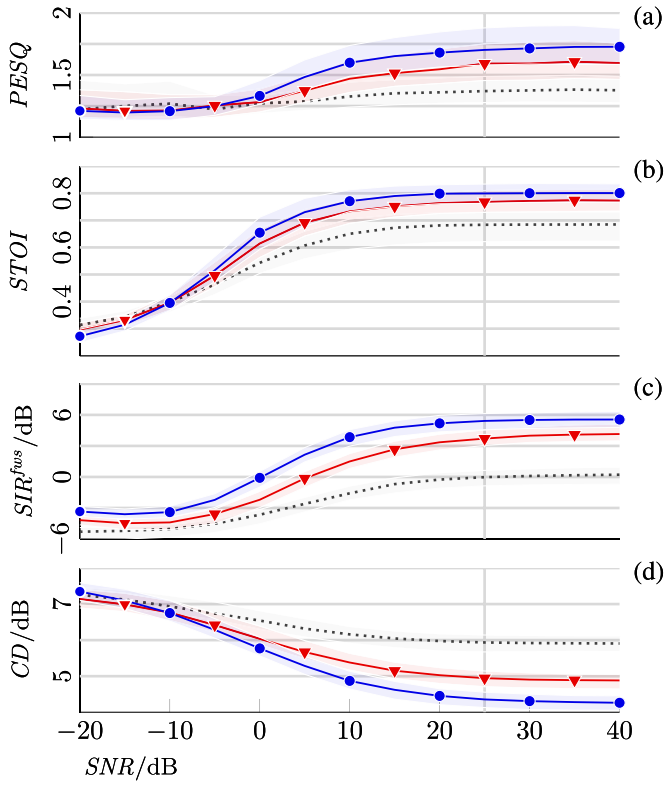


Fig. 6. (a) $PESQ$, (b) $STOI$, (c) SIR^{fws}, and (d) CD versus SNR for the reference microphone signal [·····], the MCLP+GSC Kalman filter cascade [—▼—] and the ISCLP Kalman filter [—●—] with $L = 6$ if interfering speech is present.

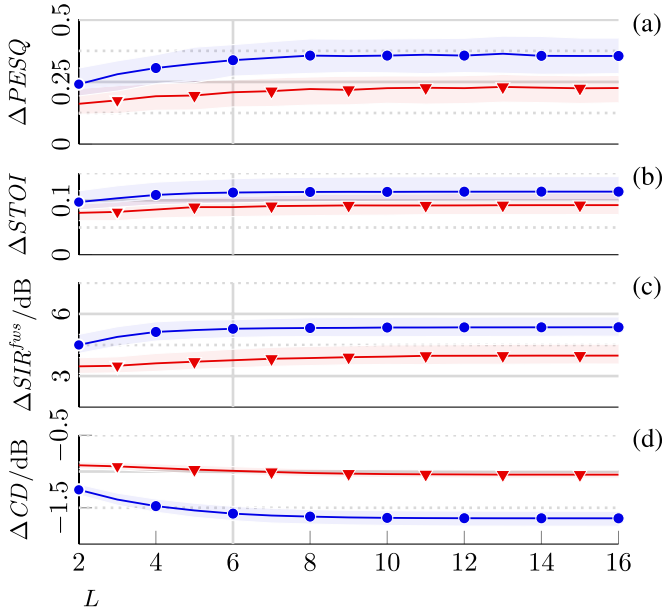


Fig. 7. (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the MCLP+GSC Kalman filter cascade [—▼—] and the ISCLP Kalman filter [—●—] at $SNR = 25$ dB if interfering speech is present.

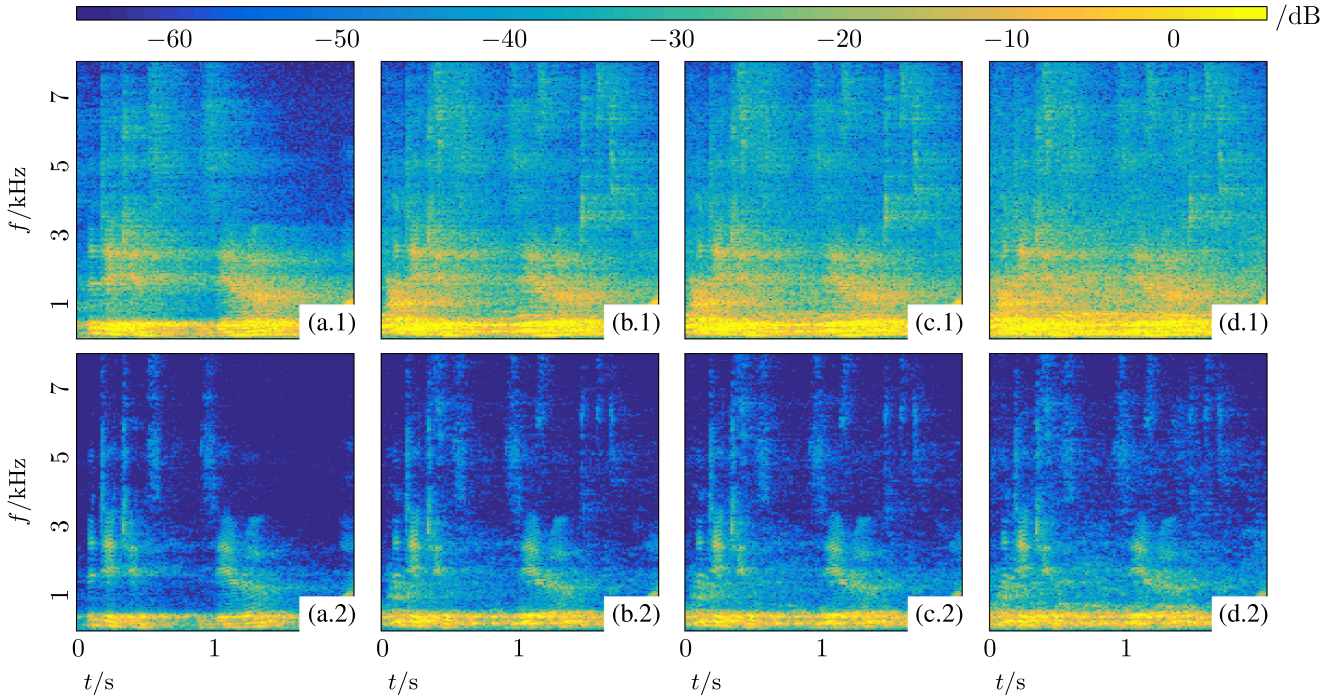


Fig. 8. Exemplary spectrograms depicting 2 s of (a.1)–(d.1) the recorded reference microphone signal and (a.2)–(d.2) the output of the ISCLP Kalman filter for $L = 16$ if interfering speech and noise are (a) absent and (b)–(d) present, at (b) $SNR = 25$ dB, (c) $SNR = 10$ dB, and (d) $SNR = 0$ dB.

Kalman filters [−■−], and the ISCLP Kalman filter [−●−] with $L = 6$ at $SNR = 10$ dB. Again, the measures largely agree. We find that after initialization, all algorithms converge after roughly 4 s. The speech source position changes at 8 s, cf. Section V-C1, such that the three algorithms have to re-adapt. In case of the ISCLP Kalman filter, this does not only require adaptation of $\hat{\mathbf{w}}(l)$, but also of the estimate $\hat{\mathbf{H}}_T(l)$, cf. (18), (21), and Section IV-B. Note that none of the three algorithms is re-initialized after $t = 8$ s, but re-adapt themselves, cf. also Section IV-B for the ISCLP Kalman filter. However, we find that for all three algorithms, convergence speed after the speech source position change is somewhat reduced as compared to the initial convergence stage.

2) Case B: Fig. 6 shows the performance in terms of (a) $PESQ$, (b) $STOI$, (c) SIR^{fws}, and (d) CD versus SNR for the reference microphone signal [·····], the MCLP+GSC Kalman filter cascade [−▾−] and the ISCLP Kalman filter [−●−] with $L = 6$. Also here, the measures show a high degree of agreement. As in case A, cf. Fig. 3, the reference microphone signal reaches better scores at higher SNR values in all measures. The curves are, however, generally flatter as compared to those in Fig. 3, which is due to the now additional interfering speech component $\mathbf{x}_2(l)$, cf. Section V-C. Above roughly $SNR = -5$ dB, both algorithms show a significant improvement over the reference microphone signal in all measures, with the ISCLP Kalman filter clearly outperforming the MCLP+GSC cascade. For the ISCLP Kalman filter, as compared to case A where $\mathbf{x}_2(l) = \mathbf{0}$, cf. Fig. 3, $PESQ$ now predicts less improvement, while $STOI$ predicts more improvement, indicating different sensitivity of both measures to the additional interfering speech component $\mathbf{x}_2(l)$.

Fig. 7 depicts the performance improvement in terms of (a) $\Delta PESQ$, (b) $\Delta STOI$, (c) ΔSIR^{fws}, and (d) ΔCD versus L with respect to the reference microphone signal for the MCLP+GSC Kalman filter cascade [−▾−] and the ISCLP Kalman filter [−●−] at $SNR = 25$ dB. Again, the ISCLP Kalman filter clearly outperforms the MCLP+GSC Kalman filter cascade in the simulated range. For the ISCLP Kalman filter, as compared to case A where $\mathbf{x}_2(l) = \mathbf{0}$, cf. Fig. 4, the improvement shows a stronger dependency on L and saturates somewhat later, indicating that longer filters are required in case of additional temporally correlated components such as $\mathbf{x}_2(l)$, which is in line with the findings in [21]. As in case A, further simulations showed that for both algorithms, the dependency on L decreases with decreasing SNR values.

VI. An Example on Actual Recordings

In this section, instead of synthesizing the microphone signals from measured RIRs and artificially generated diffuse babble noise as in Section V, we present an example of the ISCLP Kalman filter applied to actual recordings. Note that for this case, an objective performance evaluation using the intrusive measures described in Section V-B is not possible, since a clean reference signal, which should contain early target speech only, cannot be observed separately. Hence, we refer the interested reader to the corresponding audio examples [25] and limit the evaluation to a qualitative discussion of spectrograms.

Recordings were performed in a lab of the department of electrical engineering (ESAT) at KU Leuven. The selected room exhibits a comparably high reverberation time of 1.5 s and a critical distance of 0.45 m for omnidirectional sources. Similar

to Section V-B, we use a linear microphone array of $M = 5$ microphones with 8 cm inter-microphone distance. Two loudspeakers resembling a target and an interfering source are placed in 2 m distance (i.e., at more than four times the critical distance) at 0° and 45° relative to broadside direction, emitting male and female speech [47], respectively. Diffuse babble noise is generated by means of eight additional loudspeakers, placed around the setup in an arbitrary manner. Each of these loudspeakers emits randomly selected speech segments [49], with three speech segments overlaid at a time. We use the same algorithmic settings as described in Section V-D except for the following parameters. In order to account for the larger reverberation time as compared to the simulations in Section V, we now set $L = 16$ in (23) and $10 \log_{10} \bar{\psi}_{w_{LP}} = -2$ dB in (54).

Fig. 8 shows examplary spectrograms of 2 s of (a.1)–(d.1) the recorded reference microphone signal and (a.2)–(d.2) the output of the ISCLP Kalman filter for different acoustic conditions. The recording in Fig. 8(a.1) contains reverberant target speech only. As can be seen in Fig. 8(a.2), late reverberation is effectively suppressed by ISCLP Kalman filter. The recording in Fig. 8(b.2) contains both reverberant target and interfering speech as well as diffuse babble noise at $SNR = 25$ dB. The interfering speech component can be seen, e.g., at around 1.5 s. In Fig. 8(b.2), residual early interfering speech is most prominently observable around 6.01 kHz, where spatial aliasing occurs for the given setup and hence the early target and interfering source images become spatially indistinguishable. In Fig. (c)–(d), the experiment of Fig. (b) is repeated for $SNR = 10$ dB and $SNR = 0$ dB, respectively. As expected, further residual noise components appear in the output with increasing noise power at the input. Nevertheless, the early target speech component remains predominant. Informal listening tests confirm increased quality and intelligibility of the enhanced signals for all presented acoustic conditions.

VII. Conclusion

In this paper, in order to jointly perform deconvolution and spatial filtering, allowing for dereverberation, interfering speech cancellation and noise reduction, we have presented the ISCLP Kalman filter, which integrates MCLP and the GSC. Hereat, the SC filter and the LP filter operate in parallel but on different input-data frames, and are estimated jointly. We further have proposed a spectral Wiener gain post-processor, relating to the Kalman filter's posterior state estimate. Implementational aspects such as spatio-temporal target component leakage, target PSD estimation and RETF updates, as well as process equation parameter tuning and initialization have been discussed. The presented ISCLP Kalman filter has been benchmarked in terms of its dependency on the SNR and the filter length L, as well as in terms of its convergence behavior. With M the number of microphones, the ISCLP Kalman filter is roughly M^2 times less expensive than both reference algorithms, namely first a pair of alternating Kalman filters in an original and a modified version, and second an MCLP+GSC Kalman filter cascade. Nonetheless, simulation results indicate better or similar performance as compared to the original or modified version of the former, and better performance as compared to the latter.

References

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[2] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 945–958, May 2013.

[3] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 240–251, Feb. 2015.

[4] O. Schwartz, S. Gannot, and E. A. P. Habets, "Nested generalized sidelobe canceller for joint dereverberation and noise reduction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, USA, Apr. 2015, pp. 106–110.

[5] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, Jan. 2007.

[6] T. Nakatani, B. H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.

[7] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 3733–3736.

[8] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[9] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Jul. 2012.

[10] T. Yoshioka, "Dereverberation for reverberation-robust microphone arrays," in *Proc. 21st Eur. Signal Process. Conf.*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[11] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, Jun. 2015.

[12] M. Delcroix *et al.*, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2015, pp. 1–15, Dec. 2015.

[13] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech Recognit., Understanding*, Scottsdale, AZ, USA, Dec. 2015, pp. 436–443.

[14] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Partitioned block frequency domain Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Xi'An, China, Sep. 2016, pp. 1–5.

[15] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1741–1745, Dec. 2016.

[16] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

[17] T. Dietzen, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Waterschoot, "Low complexity Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, New Paltz, NY, USA, Oct. 2017.

[18] S. Braun and E. A. P. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 240–251, Jun. 2018.

[19] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Tokyo, Japan, Sep. 2018, pp. 466–470.

[20] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Tokyo, Japan, Sep. 2018, pp. 221–225.

[21] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, Mar. 2019.

[22] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, Jun. 2019.
[23] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 1561–1565.
[24] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, to be published, doi: 10.1109/TASLP.2020.2966891.
[25] T. Dietzen, "GitHub repository: Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," Jul. 2019. [Online]. Available: https://github.com/tdietzen/ISCLP-KF
[26] S. Braun *et al.*, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.
[27] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1102–1114, Jun. 2018.
[28] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. Hoboken, NJ, USA: Wiley, 2010, pp. 269–302.
[29] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
[30] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, May 2005.
[31] D. Simon, *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Hoboken, NJ, USA: Wiley, 2006.
[32] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, 1st ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[33] S. Haykin, *Adaptive Filter Theory*, 4th ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.
[34] 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, 2006.
[35] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.
[36] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, Apr. 2007.
[37] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.
[38] J. Scheuing, and B. Yang, "Correlation-based tdoa-estimation for multiple sources in reverberant environments," in *Speech and Audio Processing in Adverse Environments*. Berlin, Germany: Springer, 2008, pp. 381–416.
[39] Z. Chen, G. Gokeda, and Y. Yu, *Introduction to Direction-of-Arrival Estimation*. Norwood, MA, USA: Artech House, 2010.
[40] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 204–210, Jul. 2000.
[41] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Process.*, vol. 15, pp. 43–56, Jul. 1988.
[42] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.
[43] ITU-T, "Perceptual evaluation of of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," in *ITU-T Recommendation P.862, Int. Telecommun. Union*, Geneva, Switzerland, Feb. 2001.
[44] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[45] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[46] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Antibes, Juan les Pins, France, Sep. 2014, pp. 313–317.
[47] Bang and Olufsen, "Music for Archimedes," Compact Disc B&O, 1992.
[48] Auditec, "Auditory tests (revised)," Compact Disc Auditec, 1997.
[49] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.



Thomas Dietzen received the Dipl.-Ing. degree from Kaiserslautern University, Kaiserslautern, Germany, in 2011, and the Ph.D. degree from KU Leuven, Leuven, Belgium, in 2019, where he currently holds a Postdoctoral Research position.

Between 2012 and 2014, he was a Research Assistant with the University of Heidelberg, Heidelberg, Germany, and Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany. From 2014 to 2017, he was a Doctoral Researcher with NXP Semiconductors Belgium NV, Belgium. His research is focused on room acoustic modeling and signal enhancement in adverse acoustic conditions, specifically on spatio-temporal adaptive filtering and power-spectral-density estimation. He was a Reviewer for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, and *EURASIP Journal on Audio, Speech, and Music Processing*.



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, Leuven, Belgium, in 1997 and 2003. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Cognitive Systems Laboratory (McMaster University, 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 Division Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany. His research activities center around 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 is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing, and the EAA Technical Committee on Audio Signal Processing. He was and is involved in several large-scale national and European research projects (ITN DREAMS, Cluster of Excellence Hearing4all, and CRC Hearing Acoustics). He was the Technical Program Chair for the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in 2013 and the Chair of the ITG Conference on Speech Communication in 2018. In addition, he was a Guest Editor for several special issues (IEEE SIGNAL PROCESSING MAGAZINE and Elsevier's *Signal Processing*) and was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and *EURASIP Journal on Advances in Signal Processing*. He 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, and VDE Information Technology Society 2019).



Marc Moonen (Fellow, IEEE) is currently a Full Professor with the Electrical Engineering Department, KU Leuven, Leuven, Belgium, where he is also heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL, and audio signal processing.

Mr. Moonen is a Fellow of EURASIP. He was the Chairman for the IEEE Benelux Signal Processing Chapter, from 1998 to 2002, a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and the President of EURASIP (European Association for Signal Processing, from 2007 to 2008 and from 2011 to 2012). He was the Editor-in-Chief for the *EURASIP Journal on Applied Signal Processing*, from 2003 to 2005, an Area Editor for feature articles in the IEEE SIGNAL PROCESSING MAGAZINE, from 2012 to 2014, and has been a member of the Editorial Board of Signal Processing, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II, IEEE SIGNAL PROCESSING MAGAZINE, *Integration—The VLSI Journal*, *EURASIP Journal on Wireless Communications and Networking*, and *EURASIP Journal on Advances in Signal Processing*. He was the recipient of the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He was also the recipient of the journal best paper awards from the IEEE TRANSACTIONS ON SIGNAL PROCESSING (with Geert Leus and with Daniele Giacobello) and from Elsevier's *Signal Processing* (with Simon Doclo).



Toon van Waterschoot (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering in 2001 and 2009, respectively, from KU Leuven, Leuven, Belgium, where he is currently an Associate Professor and Consolidator Grantee of the European Research Council.

He has previously also held teaching and research positions with Delft University of Technology, Delft, The Netherlands, and the University of Lugano, Lugano, Switzerland. His research interests are in signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction. He is a member of EURASIP, ASA, and AES. He has been an Associate Editor for the *Journal of the Audio Engineering Society* and for the *EURASIP Journal on Audio, Music, and Speech Processing*, and as a Guest Editor for Elsevier's *Signal Processing*. He is the Director of the European Association for Signal Processing (EURASIP), a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, a Member of the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and a Founding Member of the EAA Technical Committee in Audio Signal Processing. He was the General Chair for the 60th AES International Conference, Leuven, Belgium in 2016, and has been on the Organizing Committee of the European Conference on Computational Optimization (EUCCO 2016), the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017), and the 28th European Signal Processing Conference (EUSIPCO 2020).