JOINT MULTI-MICROPHONE SPEECH DEREVERBERATION AND NOISE REDUCTION USING INTEGRATED SIDELOBE CANCELLATION AND LINEAR PREDICTION

Thomas Dietzen[1,2]*, Simon Doclo*[3]*, Marc Moonen*[1]*, Toon van Waterschoot*[1,2]

[1] KU Leuven, Dept. of Electrical Engineering, ESAT-STADIUS, Leuven, Belgium
[2] KU Leuven, Dept. of Electrical Engineering, ESAT-ETC, Leuven, Belgium
[3] University of Oldenburg, Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, Oldenburg, Germany

ABSTRACT

In multi-microphone speech enhancement, reverberation and noise are commonly suppressed by deconvolution and spatial filtering, i.e. 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 paper, in order to perform both deconvolution and spatial filtering, we propose to integrate MCLP and the GSC into a novel framework referred to as integrated sidelobe cancellation and linear prediction (ISCLP), wherein the sidelobe-cancellation (SC) filter and the linear prediction (LP) filter operate in parallel. Further, within this framework, we propose to estimate both filters jointly by means of a single Kalman filter. While ISCLP is roughly M times less expensive than a corresponding cascade of multiple-output MCLP and the GSC, where M denotes the number of microphones, it performs equally well in terms of dereverberation and noise reduction, as shown in simulations using one localized noise source.

***Index Terms*—** Dereverberation, Noise Reduction, Beamforming, Multi-Channel Linear Prediction, Kalman Filter, Generalized Eigenvalue Decomposition

1. INTRODUCTION

In many wide-spread speech processing applications such as hands-free telephony and distant automatic speech recognition, reverberation and additive noise impinging on a microphone may deteriorate the quality and intelligibility of the speech recordings. The demanding tasks of dereverberation, noise reduction, and in particular the conjunction of both therefore remain a subject of ongoing research, with multi-microphone-based approaches exploiting spatial diversity receiving particular interest [1–13].

As a spatial filtering technique, beamforming is commonly used in noise reduction, but may as well be applied for dereverberation [1–3]. In order to perform both dereverberation and noise reduction, several beamforming schemes have been proposed. In [1], a cascaded approach is presented, using data-independent, super-directive 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 [2, 3]. In [2], joint dereverberation and noise reduction is performed using a single GSC, while in [3], 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.

As a deconvolution technique, multi-channel linear prediction (MCLP) [4–13] 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. Besides iterative LP filter estimation approaches such as [4, 5, 7, 8], also adaptive approaches based on recursive least squares [6, 11] as well as the Kalman filter [9, 10, 12] have evolved in the past years. In order to reduce noise after dereverberation, multiple-output MCLP has been cascaded with MVDR beamforming in [8]. In [13], joint MCLP-based dereverberation and noise reduction is performed using two Kalman filters, alternately estimating the LP filter and the noise-free reverberant speech component.

In this paper, instead of cascading MCLP and beamforming or relying on beamforming only, we propose to integrate MCLP and the GSC by employing an SC path and LP path in parallel, resulting in a framework we refer to as integrated sidelobe cancellation and linear prediction (ISCLP). Within this novel framework, we propose to estimate the SC and LP filters jointly by means of a single Kalman filter. Here, the spatial components MF and BM require an estimate of the relative early transfer functions (RETFs), cf. also [2], while the Kalman filter requires an estimate of the power spectral density (PSD) of the early reverberant-speech component, cf. also [9,10,12]. We estimate both by means of the generalized eigenvalue decomposition (GEVD), cf. [14–16]. As compared to a corresponding cascade of multiple-output MCLP and the GSC, the ISCLP framework is computationally roughly M times less expensive, where M denotes the number of microphones. Yet, ISCLP performs equally well in terms of dereverberation and noise reduction, as shown in simulations using one localized noise source.

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven internal funding project C2-16-00449 'Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking', KU Leuven Impulsfonds project IMP/14/037, KU Leuven Internal Funds project VES/16/032, and was supported by the European Commission under Grant Agreement no. 316969 (FP7-PEOPLE Marie Curie ITN 'Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)') and no. 773268 (H2020-ERC-CoG 'The Spatial Dynamics of Room Acoustics (SONORA)'), and by the Flemish Government under Project no. 150611 (VLAIO O&O Project 'Proof-of-concept of a Rationed Architecture for Vehicle Entertainment and NVH Next-generation Acoustics (RAVENNA)') and no. HBC.2016.0085 (VLAIO TETRA Project 'Innovative use of sensors in mobile platforms (m-sense)'). The scientific responsibility is assumed by its authors.

978-1-5386-8151-0/18/$31.00 ©2018 IEEE

2. SIGNAL MODEL

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 m^{th} microphone signal. In the following, we treat all frequency bins independently and hence omit the frequency index. We define the stacked microphone signal vector[1] $\mathbf{y}(l) \in \mathbb{C}^M$,

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

composed of the reverberant-speech component $\mathbf{x}(l)$ and the noise component $\mathbf{v}(l)$, defined similarly to (1),

$$\mathbf{y}(l) = \mathbf{x}(l) + \mathbf{v}(l). \tag{2}$$

Here, the reverberant-speech component $\mathbf{x}(l)$ may be decomposed into the early and late components $\mathbf{x}_e(l)$ and $\mathbf{x}_\ell(l)$, where the early components in $\mathbf{x}_e(l)$ are related by the (presumed time-invariant) RETFs in $\mathbf{h} \in \mathbb{C}^M$, defined relative to the early transfer function of the first microphone, i.e.

$$\mathbf{x}(l) = \mathbf{x}_e(l) + \mathbf{x}_\ell(l), \tag{3}$$

$$\mathbf{x}_e(l) = x_e(l)\mathbf{h}, \tag{4}$$

$$\mathbf{h} = \big(1 \ \ h_2 \ \cdots \ h_M\big)^T = \big(1 \ \ \mathbf{h}_{2:M}^T\big)^T. \tag{5}$$

In the following, we assume that $x_e(l)$ is temporally uncorrelated, i.e. $E\big\{x_e(l-l')x_e^*(l)\big\} = 0$ for $l' \neq 0$. For speech signals, this assumption can be considered justified if the frame length and frame shift are sufficiently large. Further, we assume that $\mathbf{x}_e(l)$, $\mathbf{x}_\ell(l)$, and $\mathbf{v}(l)$ are mutually uncorrelated within frame l, and that $\mathbf{x}_\ell(l)$ may be modeled as a diffuse component with coherence matrix $\mathbf{\Gamma} \in \mathbb{C}^{M\times M}$. Let $\mathbf{\Psi}_y(l) = E\big\{\mathbf{y}(l)\mathbf{y}^H(l)\big\} \in \mathbb{C}^{M\times M}$ denote the microphone signal correlation matrix, and let $\mathbf{\Psi}_x(l)$ and $\mathbf{\Psi}_v(l)$ be defined similarly. With (2)–(4), we then find

$$\begin{aligned}\mathbf{\Psi}_y(l) &= \mathbf{\Psi}_x(l) + \mathbf{\Psi}_v(l)\\ &= \psi_{x_e}(l)\mathbf{h}\mathbf{h}^H + \psi_{x_\ell}(l)\mathbf{\Gamma} + \mathbf{\Psi}_v(l),\end{aligned} \tag{6}$$

with $\psi_{x_e}(l)$ and $\psi_{x_\ell}(l)$ the power spectral densities (PSDs) of the early and late reverberant-speech components, respectively. The diffuse coherence matrix $\mathbf{\Gamma}$ may be computed from the microphone array geometry [15, 16].

In this paper, although the presented ISCLP framework is not restricted to this scenario, we evaluate the case where $\mathbf{v}(l)$ originates from a single localized noise source, cf. Sec. 5, i.e. $\mathbf{v}(l)$ may be decomposed in a similar manner as $\mathbf{x}(l)$.

3. INTEGRATED SIDELOBE CANCELLATION AND LINEAR PREDICTION

We strive to estimate the early reverberant-speech component $x_e(l)$ from the microphone signals $\mathbf{y}(l)$ defined in Sec. 2. For this purpose, we introduce the ISCLP framework. In Sec. 3.1, we describe the SC and LP filter constellation, which requires spatio-temporal pre-processing of $\mathbf{y}(l)$. In Sec. 3.2, we discuss a recursive filter estimation procedure, which is based on a single Kalman filter.

[1]Notation: vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, with zero and identity matrices denoted by $\mathbf{0}$ and $\mathbf{I}$, respectively. The notations $\circ^T$, $\circ^*$, $\circ^H$, $E\{\circ\}$, and $\hat{\circ}$ denote the transpose, the complex conjugate, the complex conjugate transpose, the expected value, and the estimate of a matrix $\circ$, respectively.

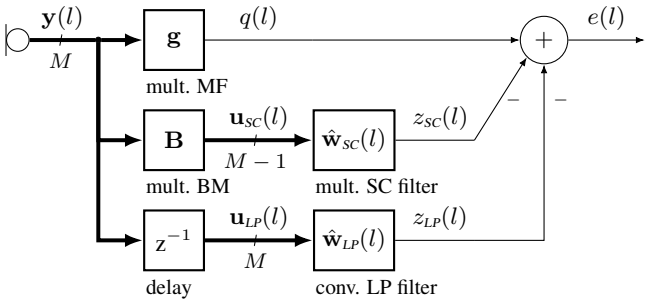

Fig. 1. The integrated sidelobe cancellation and linear prediction (ISCLP) framework.

3.1. ISCLP Signal Paths

The ISCLP framework depicted in Fig. 1 integrates the GSC and MCLP frameworks 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 linear-prediction path, composed of a delay and an LP filter. While the MF, the BM and the SC filter are multiplicative, the LP filter is convolutive. Structurally, one may interpret the ISCLP framework either as MCLP with the traditional reference channel selection replaced by a GSC, or alternatively as a GSC employing a generalized BM (composed of a traditional BM and a delay), and a convolutive filter (composed of the SC and the LP filter).

The ideal MF $\mathbf{g} \in \mathbb{C}^M$ is given by

$$\mathbf{g} = \mathbf{h}/\|\mathbf{h}\|^2, \tag{7}$$

requiring an estimate of $\mathbf{h}$ in practice, which we obtain as shown in Sec. 4. For the MF output $q(l)$, combining (2)–(4), we then find

$$\begin{aligned}q(l) &= \mathbf{g}^H\mathbf{y}(l)\\ &= x_e(l) + \underbrace{\mathbf{g}^H\mathbf{x}_\ell(l)}_{q_{x_\ell}(l)} + \underbrace{\mathbf{g}^H\mathbf{v}(l)}_{q_v(l)}.\end{aligned} \tag{8}$$

Per definition, the ideal BM $\mathbf{B} \in \mathbb{C}^{M\times M-1}$ is orthogonal to $\mathbf{g}$, i.e. $\mathbf{B}^H\mathbf{g} = \mathbf{0}$ and hence $\mathbf{B}^H\mathbf{h} = \mathbf{0}$, which may be implemented as

$$\mathbf{B} = \big(-\mathbf{h}_{2:M} \ \ \mathbf{I}\big)^H. \tag{9}$$

The SC-filter input $\mathbf{u}_{SC}(l) \in \mathbb{C}^{M-1}$ is then given by

$$\begin{aligned}\mathbf{u}_{SC}(l) &= \mathbf{B}^H\mathbf{y}(l)\\ &= \mathbf{B}^H\mathbf{x}_\ell(l) + \mathbf{B}^H\mathbf{v}(l),\end{aligned} \tag{10}$$

whereby the early reverberant-speech component $\mathbf{x}_e(l) = x_e(l)\mathbf{h}$ is canceled. Using a delay of one 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.

$$\mathbf{u}_{LP}(l) = \big(\mathbf{y}^T(l-1) \ \cdots \ \mathbf{y}^T(l-L+1)\big)^T. \tag{11}$$

Note that due to the delay, $\mathbf{u}_{LP}(l)$ is uncorrelated to $x_e(l)$ if $x_e(l)$ itself is temporally uncorrelated. With the SC filter $\hat{\mathbf{w}}_{SC}(l) \in \mathbb{C}^{M-1}$ and the LP filter $\hat{\mathbf{w}}_{LP}(l) \in \mathbb{C}^{(L-1)M}$, the enhanced signal at the output of the ISCLP framework is given by

$$e(l) = q(l) - \underbrace{\hat{\mathbf{w}}_{SC}^H(l)\mathbf{u}_{SC}(l)}_{z_{SC}(l)} - \underbrace{\hat{\mathbf{w}}_{LP}^H(l)\mathbf{u}_{LP}(l)}_{z_{LP}(l)}. \tag{12}$$

For $\hat{\mathbf{w}}_{SC}^H(l)$ and $\hat{\mathbf{w}}_{LP}^H(l)$, we seek a set of filters that ideally yields $e(l) = x_e(l)$, which requires $z_{SC}(l)+z_{LP}(l) = q_{x_\ell}(l)+q_v(l)$, cf. (8). Note that $\mathbf{u}_{SC}(l)$ in (10) depends on the current frame of $\mathbf{y}(l)$ only, such that $\hat{\mathbf{w}}_{SC}(l)$ will exploit spatial correlations within the current frame, while $\mathbf{u}_{LP}(l)$ in (11) depends on the $L-1$ previous frames of $\mathbf{y}(l)$, 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). Since both $q_{x_\ell}(l)$ and $q_v(l)$ may exhibit spatial and spatio-temporal correlations within and across frames, we do not restrict the SC and LP filter paths to suppress only either of the two components each, but instead they may *jointly* suppress both components. Therefore, we strive to estimate both filters jointly.

3.2. ISCLP Kalman Filter

In order to recursively estimate the SC and LP filter, we employ a Kalman filter, which has also been applied successfully to MCLP in previous works [9,10,12]. Hereby, we interpret $\hat{\mathbf{w}}_{SC}(l)$ and $\hat{\mathbf{w}}_{LP}(l)$ as estimates of the hidden states $\mathbf{w}_{SC}(l)$ and $\mathbf{w}_{LP}(l)$ leading to complete cancellation of $q_{x_\ell}(l)+q_v(l)$, and therefore yielding $e(l) = x_e(l)$. In the following, we first define the state equations, comprising the so-called observation equation and process equation, and then present the corresponding Kalman filter update equations.

We stack the SC and LP filter path into $\mathbf{u}(l) \in \mathbb{C}^{LM-1}$ and $\mathbf{w}(l) \in \mathbb{C}^{LM-1}$, i.e.

$$\mathbf{u}(l) = \left(\mathbf{u}_{SC}^T(l) \;\; \mathbf{u}_{LP}^T(l)\right)^T, \tag{13}$$

$$\mathbf{w}(l) = \left(\mathbf{w}_{SC}^T(l) \;\; \mathbf{w}_{LP}^T(l)\right)^T. \tag{14}$$

Reformulating (12) using (13)–(14), inserting $e(l) = x_e(l)$ and rearranging yields the so-called observation equation,

$$q^*(l) = \mathbf{u}^H(l)\mathbf{w}(l) + x_e^*(l). \tag{15}$$

In Kalman filter terminology, we refer to $q^*(l)$ as the observable and to $x_e^*(l)$ as the (presumed zero-mean and temporally uncorrelated) observation noise with PSD $\psi_{x_e}(l)$ as defined in (6). In practice, in order to implement the Kalman filter update equations, an estimate of $\psi_{x_e}(l)$ is required, which we obtain as shown in Sec. 4. The so-called process equation models the evolution of the hidden state $\mathbf{w}(l)$ in the form of a first-order difference equation, i.e.

$$\mathbf{w}(l) = \mathbf{A}^H(l)\mathbf{w}(l-1) + \mathbf{w}_\Delta(l). \tag{16}$$

where $\mathbf{A}(l) \in \mathbb{C}^{(LM-1)\times(LM-1)}$ models the state transition from one frame to the next, and the process noise $\mathbf{w}_\Delta(l)$ models a random (zero-mean and temporally uncorrelated) variation component with correlation matrix $\mathbf{\Psi}_{w_\Delta}(l) \in \mathbb{C}^{(LM-1)\times(LM-1)}$. Both $\mathbf{A}(l)$ and $\mathbf{\Psi}_{w_\Delta}(l)$ are commonly considered design parameters and thereby chosen to be diagonal, with the diagonal elements of $\mathbf{A}(l)$ acting as forgetting factors.

The hidden state $\mathbf{w}(l)$ modeled by (15)–(16) may be estimated recursively by means of the Kalman filter update equations [17],

$$\hat{\mathbf{w}}(l) = \mathbf{A}^H(l)\hat{\mathbf{w}}^+(l-1), \tag{17}$$

$$\mathbf{\Psi}_w(l) = \mathbf{A}^H(l)\mathbf{\Psi}_w^+(l-1)\mathbf{A}(l) + \mathbf{\Psi}_{w_\Delta}(l), \tag{18}$$

$$e^*(l) = q^*(l) - \mathbf{u}^H(l)\hat{\mathbf{w}}(l), \tag{19}$$

$$\psi_e(l) = \mathbf{u}^H(l)\mathbf{\Psi}_w(l)\mathbf{u}(l) + \psi_{x_e}(l), \tag{20}$$

$$\mathbf{k}(l) = \mathbf{\Psi}_w(l)\mathbf{u}(l)\psi_e^{-1}(l), \tag{21}$$

$$\hat{\mathbf{w}}^+(l) = \hat{\mathbf{w}}(l) + \mathbf{k}(l)e^*(l), \tag{22}$$

$$\mathbf{\Psi}_w^+(l) = \mathbf{\Psi}_w(l) - \mathbf{k}(l)\mathbf{u}^H(l)\mathbf{\Psi}_w(l), \tag{23}$$

initialized by $\hat{\mathbf{w}}^+(0)$ and $\mathbf{\Psi}_w^+(0)$. Here, (17)–(18) and (22)–(23) are respectively referred to as the time update and the measurement update of the state estimate $\hat{\mathbf{w}}(l)$ and the state estimation error correlation matrix $\mathbf{\Psi}_w(l) \in \mathbb{C}^{(LM-1)\times(LM-1)}$, where the superscript $^+$ provides notational distinction. While the time update reflects the state evolution, cf. (16), the measurement update processes the current observation, cf. (15), by incorporating (19)–(21), which in turn yield the enhanced signal $e^*(l)$, its PSD $\psi_e(l)$, and the so-called Kalman gain vector $\mathbf{k}(l)$. The enhanced signal $e^*(l)$ in (19) thereby represents the Kalman filter estimate of $x_e^*(l)$. During convergence, the norm of $\mathbf{\Psi}_w(l)$ decreases, such that $e^*(l)$ and $\psi_e(l)$ in (19)–(20) converge to $x_e^*(l)$ and $\psi_{x_e}(l)$, respectively.

4. RETF AND PSD ESTIMATION

As apparent from (7), (9), and (20), the ISCLP Kalman filter requires an estimate of the RETF $\mathbf{h}$ and the early reverberant-speech PSD $\psi_{x_e}(l)$. As previously proposed in [14–16], we obtain $\hat{\mathbf{h}}$ as well as an estimate $\hat{\psi}_{x_\ell}(l)$ of the late reverberant-speech PSD by means of the GEVD. Based on $\hat{\psi}_{x_\ell}(l)$, we then obtain $\hat{\psi}_{x_e}(l)$ as outlined in the following. Within the limits of this paper, we assume that an estimate $\hat{\mathbf{\Psi}}_v(l)$ of the noise correlation matrix is available, e.g. estimated from noise-only frames if $\mathbf{v}(l)$ is stationary. With $\hat{\mathbf{\Psi}}_y(l)$ recursively computed from $\mathbf{y}(l)$, i.e.

$$\hat{\mathbf{\Psi}}_y(l) = \beta\hat{\mathbf{\Psi}}_y(l-1) + (1-\beta)\mathbf{y}(l)\mathbf{y}^H(l), \tag{24}$$

where β is a forgetting factor, we obtain the estimate $\hat{\mathbf{\Psi}}_x(l) = \hat{\mathbf{\Psi}}_y(l) - \hat{\mathbf{\Psi}}_v(l)$ according to (6). In each frame l, we then perform the GEVD of $\hat{\mathbf{\Psi}}_x(l)$ and $\mathbf{\Gamma}$, defined by

$$\hat{\mathbf{\Psi}}_x(l)\mathbf{P} = \mathbf{\Gamma}\mathbf{P}\mathbf{\Lambda}(l), \tag{25}$$

where the diagonal elements of $\mathbf{\Lambda}(l) \in \mathbb{R}^{M\times M}$ comprise the generalized eigenvalues $\lambda_m(l)$ with $\lambda_1(l)$ the maximum generalized eigenvalue, and the columns of $\mathbf{P} \in \mathbb{C}^{M\times M}$ comprise the corresponding generalized eigenvectors $\mathbf{p}_m$. The eigenvectors are uniquely defined up to a scaling factor, and we scale them such that

$$\mathbf{P}^H\mathbf{\Gamma}\mathbf{P} = \mathbf{I}, \tag{26}$$

and hence $\mathbf{P}^H\hat{\mathbf{\Psi}}_x(l)\mathbf{P} = \mathbf{\Lambda}(l)$. If $\hat{\mathbf{\Psi}}_x(l)$ obeys the model in (6), inserting in (25) yields

$$\mathbf{\Gamma}\mathbf{P}\mathbf{\Lambda}(l) = \hat{\psi}_{x_e}(l)\hat{\mathbf{h}}\hat{\mathbf{h}}^H\mathbf{P} + \hat{\psi}_{x_\ell}(l)\mathbf{\Gamma}\mathbf{P}, \tag{27}$$

or equivalently, left-multiplying by $\mathbf{P}^H$ while using (26),

$$\mathbf{\Lambda}(l) = \hat{\psi}_{x_e}(l)\mathbf{P}^H\hat{\mathbf{h}}\hat{\mathbf{h}}^H\mathbf{P} + \hat{\psi}_{x_\ell}(l)\mathbf{I}. \tag{28}$$

Since $\mathbf{P}^H\hat{\mathbf{h}}\hat{\mathbf{h}}^H\mathbf{P}$ in (28) is a diagonal rank-1 matrix, only the first diagonal element $\mathbf{p}_1^H\hat{\mathbf{h}}\hat{\mathbf{h}}^H\mathbf{p}_1$ is different from zero. The eigenvalues $\lambda_m(l)$ may therefore be written as

$$\lambda_m(l) = \begin{cases} \hat{\psi}_{x_e}(l)\mathbf{p}_m^H\hat{\mathbf{h}}\hat{\mathbf{h}}^H\mathbf{p}_m + \hat{\psi}_{x_\ell}(l) & \text{for } m = 1, \\ \hat{\psi}_{x_\ell}(l) & \text{else.} \end{cases} \tag{29}$$

Considering the first eigenvalue-eigenvector pair in (27) and rearranging yields $\hat{\mathbf{h}} = \left(\lambda_1(l) - \hat{\psi}_{x_\ell}(l)\right)\mathbf{\Gamma}\mathbf{p}_1 / \left(\hat{\psi}_{x_e}(l)\hat{\mathbf{h}}^H\mathbf{p}_1\right)$, i.e. $\hat{\mathbf{h}}$ is proportional to $\mathbf{\Gamma}\mathbf{p}_1$, cf. [14]. Now that the first element of $\mathbf{h}$ equals one by definition, cf. (5), we can estimate $\mathbf{h}$ as

$$\hat{\mathbf{h}} = \mathbf{\Gamma}\mathbf{p}_1 / (\mathbf{i}_1^T\mathbf{\Gamma}\mathbf{p}_1), \tag{30}$$

where $\mathbf{i}_1$ denotes the first column of $\mathbf{I}$. Inserting (30) into (29) and noting that $\mathbf{p}_1^H \mathbf{\Gamma} \mathbf{p}_1 = 1$, cf. (26), we can estimate $\psi_{x_e}(l)$ as

$$\hat{\psi}_{x_e}(l) = |\mathbf{i}_1^T \mathbf{\Gamma} \mathbf{p}_1|^2 \big(\lambda_1(l) - \hat{\psi}_{x_\ell}(l)\big), \tag{31}$$

where in theory $\hat{\psi}_{x_\ell}(l) = \lambda_{m\neq 1}(l)$, cf. (29) and [15, 16]. Note that in practice, due to modeling and estimation errors, $\hat{\mathbf{\Psi}}_x(l)$ does not perfectly obey (6), such that the individual eigenvalues $\lambda_{m\neq 1}(l)$ will differ to some extent. Hence, we may alternatively obtain $\hat{\psi}_{x_\ell}(l)$ by averaging over $\lambda_{m\neq 1}(l)$ [15, 16]. Note that the estimator in (31) has not been proposed previously, instead however, one may obtain $\hat{\psi}_{x_e}(l)$ from $\hat{\psi}_{x_\ell}(l)$ using the decision-directed approach [12, 15, 16]. Further, for the same reason, $\mathbf{P}$ and hence $\hat{\mathbf{h}}$ will typically not be perfectly time-invariant. In order to achieve a time-invariant estimate $\hat{\mathbf{h}}$, one may average over a number of selected frames.

5. SIMULATIONS

In our simulations, we compare ISCLP with multiple-output MCLP cascaded by a GSC, subsequently referred to as MCLP+GSC. In MCLP+GSC, we estimate the LP and SC filters independently. Conceptually, this cascade relates to the MCLP+MVDR cascade presented in [8], where the LP filters are estimated using the (iterative) weighted prediction error (WPE) method [5]. For the sake of a meaningful comparison however, instead of using WPE, the estimation of the LP and SC filters in MCLP+GSC has been implemented in a similar manner as proposed for ISCLP, i.e. using Kalman filtering and the GEVD, cf. Sec. 3 and Sec. 4. The (convolutive) MCLP component, here creating one output signal per microphone, requires M Kalman filters with state length $(L-1)M$ each. The (multiplicative) GSC component, applied to the MCLP outputs, requires one Kalman filter with state length $M-1$. Note that ISCLP requires only a single Kalman filter with state length $LM-1$, cf. Sec. 3.2, and is therefore computationally roughly M times less expensive.

We use RIRs of $M = 5$ microphones with 8 cm spacing measured in a room with 610 ms reverberation time [18]. The speech source, emitting male speech [19], is positioned at 2 m distance in the broadside direction of the microphone array. We simulate 16 realizations, each using a randomly selected 10 s long segment of the speech file. The noise component originates from a single localized source emitting (stationary) speech-shaped noise, positioned in 2 m distance with an angle of $(30,\ 60,\ 90)^\circ$ relative to the speech source. The SNR, defined as the power ratio of $x_1(l)$ and $v_1(l)$ in the time domain, ranges between -20 and 40 dB. The STFT analysis and synthesis is based on square-root Hann windows of $N_{STFT} = 512$ samples with 50% overlap at $f_s = 16$ kHz. The recursive estimate $\hat{\mathbf{\Psi}}_y(l)$ is computed using $\beta = e^{-N_{STFT}/(2f_s\tau)}$ with $\tau = 10$ ms, while the (time-invariant) estimate $\hat{\mathbf{\Psi}}_v$ is computed from 3 s noise-only frames. As opposed to $\hat{\psi}_{x_e}(l)$, computed from $\hat{\mathbf{\Psi}}_x(l) = \hat{\mathbf{\Psi}}_y(l) - \hat{\mathbf{\Psi}}_v$ in each frame, the (time-invariant) RETF estimate $\hat{\mathbf{h}}$ is computed from the average of $\hat{\mathbf{\Psi}}_x(l)$ over the entire realization. In ISCLP and MCLP+GSC, we set $L = 29$ and initialize all filters as $\hat{\mathbf{w}}^+(0) = \mathbf{0}$, while the initial state error correlation matrix $\mathbf{\Psi}_w^+(0)$ is chosen to be diagonal in all simulations. For the LP path in ISCLP and the MCLP component in MCLP+GSC, expecting lower values for later prediction coefficients, we choose the power of the corresponding diagonal elements in $\mathbf{\Psi}_w^+(0)$ to drop by 2 dB each M elements. We set the process noise correlation matrix and the state transition matrix to $\mathbf{\Psi}_{\Delta_w}(l) = \alpha \mathbf{\Psi}_w^+(0)$ and $\mathbf{A}(l) = \sqrt{1-\alpha}\,\mathbf{I}$, respectively, with $10\log_{10}\alpha = -25$ dB. For the evaluation, we compute the short-time objective intelligibility measure (STOI) [20] $\in [0, 1]$ after convergence. STOI is computed for the speech-plus-noise mixture and the speech component only, referred to as $STOI_{x+v}$ and $STOI_x$ indicating dereverberation-plus-noise-reduction and dereverberation-only performance, respectively. The direct component of $x_1(l)$ is chosen as a clean reference signal, defined from a window of 1 ms around the maximum peak of the corresponding RIR. The results are averaged over all realizations and source positions.

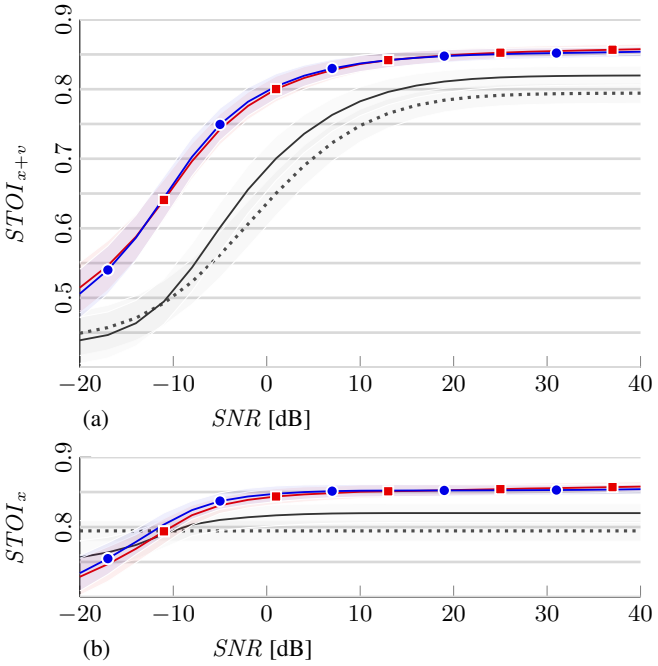

Fig. 2. (a) $STOI_{x+v}$ and (b) $STOI_x$ versus SNR for the first microphone [·····], the MF [——], MCLP+GSC [—▪—], and ISCLP [—●—]. The shaded areas represent the standard deviation.

The simulation results in terms of $STOI_{x+v}$ and $STOI_x$ are shown in Fig. 2 (a) and Fig. 2 (b), respectively. As observed from Fig. 2 (a), for the first microphone [·····], $STOI_{x+v}$ ranges between 0.45 at $SNR = -20$ dB and 0.79 at $SNR = 40$ dB, where the upper limit is determined by reverberation only, as apparent from $STOI_x$ in Fig. 2 (b). The MF [——] achieves some amount of noise reduction and dereverberation for $SNR \geq -10$ dB, reaching an improvement in $STOI_{x+v}$ of 0.05 at 0 dB and 0.025 at 40 dB, however introduces speech distortion for lower values due to stronger RETF estimation errors, i.e. scores lower than the first microphone in terms of both $STOI_{x+v}$ and $STOI_x$. MCLP+GSC [—▪—] and ISCLP [—●—] perform very similarly in both $STOI_{x+v}$ and $STOI_x$. In terms of $STOI_{x+v}$, as compared to the MF, both reach an improvement of up to 0.15 for low SNR values and 0.035 for $SNR = 40$ dB. The improvement in terms of $STOI_x$ remains constant at 0.035 for $SNR > 0$ dB and decreases for lower SNR values. Audio examples are available online [21].

6. CONCLUSION

In this paper, for the purpose of joint dereverberation and noise reduction, we have presented the ISCLP framework integrating MCLP and the GSC, wherein the SC and LP filters have been estimated jointly by means of a single Kalman filter. As compared to a corresponding MCLP+GSC cascade, while being equally performant, ISCLP is computationally roughly M times less expensive.

7. REFERENCES

[1] 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.

[2] 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, Language Proc.*, vol. 23, no. 2, pp. 240–251, Feb. 2015.

[3] 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. (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 106 – 110.

[4] 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.

[5] 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, July 2012.

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

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

[8] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, pp. 1–15, Dec. 2015.

[9] 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 Acoustic Signal Enhancement (IWAENC 2016)*, Xi'An, China, Sep. 2016, pp. 1–5.

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

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

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

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

[14] 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.

[15] I. Kodrasi and S. Doclo, "Late reverberant power spectral density estimation based on an eigenvalue decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc. (ICASSP 2017)*, New Orleans, USA, Mar. 2017, pp. 611–615.

[16] I. Kodrasi and S. Doclo, "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods," in *Proc. IEEE Hands-free Speech Com. Mic. Arrays (HSCMA 2017)*, San Francisco, CA, USA, Mar. 2017, pp. 116–120.

[17] S. Haykin, *Adaptive Filter Theory*, vol. 4th edition, Prentice-Hall, 2002.

[18] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC 2014)*, Antibes – Juan les Pins, France, Sept. 2014, pp. 313–317.

[19] Bang and Olufsen, "Music for Archimedes," Compact Disc B&O, 1992.

[20] 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.

[21] T. Dietzen, "Audio examples for IWAENC 2018," `ftp://ftp.esat.kuleuven.be/pub/SISTA/tdietzen/reports/iwaenc18/audio`, Apr. 2018.