Comparative Analysis of Generalized Sidelobe Cancellation and Multi-Channel Linear Prediction for Speech Dereverberation and Noise Reduction

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

***Abstract*—For blind speech dereverberation, two frameworks are commonly used: on the one hand, the multi-channel linear prediction (MCLP) framework, and on the other hand, data-dependent beamforming, e.g., the generalized sidelobe canceler (GSC) framework. The MCLP framework is designed to perform deconvolution and hence has gained increased prominence in blind speech dereverberation. The GSC framework is commonly used for noise reduction, but may be applied for dereverberation as well. In previous work, we have shown that for the noiseless case, MCLP and the GSC yield in theory mathematically equivalent results in terms of dereverberation. In this paper, we assume additional coherent as well as incoherent-noise components and formally analyze and compare both frameworks in terms of dereverberation and noise reduction performance. Both the theoretical analysis and time domain simulation results demonstrate that unlike the GSC, MCLP expectably shows limited performance in terms of noise reduction, while both perform equally well in terms of dereverberation, provided that the GSC blocking matrix achieves complete blocking of the early reverberant-speech component and sufficiently many microphones are available. In case of incomplete blocking, however, the GSC performs inferior to MCLP in terms of dereverberation, as shown in short-time Fourier transform domain simulations.**

***Index Terms*—Multi-channel linear prediction, data-dependent beamforming, dereverberation, noise reduction.**

Manuscript received September 6, 2018; revised December 5, 2018; accepted December 5, 2018. Date of publication December 13, 2018; date of current version December 28, 2018. This work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven internal funds C2-16-00449 "Distributed Digital Signal Processing for Ad-Hoc Wireless Local Area Audio Networking," VES/16/032, VES/19/004, and Impulse Fund IMP/14/037; IWT O&O Project 150611 "Proof-of-concept of a Rationed Architecture for Vehicle Entertainment and NVH Next-generation Acoustics (RAVENNA)"; VLAIO TETRA Project HBC.2016.0085 "m-sense: innovative use of sensors in mobile platforms" and O&O Project HBC.2017.0358 "SPOTT—Tomorrow's Scalable and PersOnalised advertising Technnology, Today"; and EU FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)," funded by the European Commission under Grant 316969. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (773268). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Richard Christian Hendriks. *(Corresponding author: Thomas Dietzen.)*

T. Dietzen was with the NXP Semiconductors Belgium NV, Leuven 3001, Belgium. He is now with the Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven 3001, Belgium (e-mail: thomas.dietzen@esat.kuleuven.be).

A. Spriet and W. Tirry are with the NXP Semiconductors Belgium NV, Leuven 3001, Belgium (e-mail: ann.spriet@nxp.com; wouter.tirry@nxp.com).

S. Doclo is with the Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4All, University of Oldenburg, Oldenburg 26111, Germany (e-mail: simon.doclo@uni-oldenburg.de).

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

Digital Object Identifier 10.1109/TASLP.2018.2886743

I. Introduction

IT IS well known that reverberation, caused by reflections against room boundaries and objects, and background noise may have a deteriorating effect on the quality and intelligibility of a speech signal recorded by a microphone [1]. Speech dereverberation accompanied by noise reduction is therefore needed in many applications ranging from hands-free mobile telephony to distant automatic speech recognition.

Dereverberation approaches based on multiple microphones take advantage of spatial diversity and, according to the multiple input/output inverse theorem (MINT) [2], theoretically allow complete inversion of the (presumed time-invariant) room impulse responses (RIRs) between the speech source and the microphone array, provided that the corresponding transfer functions do not share common zeros. In practical applications however, the RIRs are unknown – and since MINT is very sensitive to RIR estimation errors [3], which are unavoidable in practice, especially in noisy environments [4], [5], explicit inversion is not favorable. In recent years instead, assuming *no or limited prior knowledge* on the RIRs, multi-channel linear prediction (MCLP) [6]–[17], beamforming [18]–[25] and combinations thereof [26]–[28] have been most commonly and successfully used for (blind) speech dereverberation, while partly including noise reduction [17], [21], [22], [25]–[28]. In the following, we briefly review these approaches.

The MCLP framework is designed to perform deconvolution, and is hence suited for dereverberation, while noise reduction is *not* targeted. It operates blindly on the microphone signals, i.e., does not require any prior knowledge on the RIRs. A block diagram of MCLP is shown in Fig. 1. The framework relies on the premise that the reverberant component to be canceled can be modeled as a filtered version of the delayed microphone signals, i.e., as a linear prediction component. The prediction delay is a design parameter defining the number of early reflections to be maintained. The sole task in MCLP

2329-9290 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

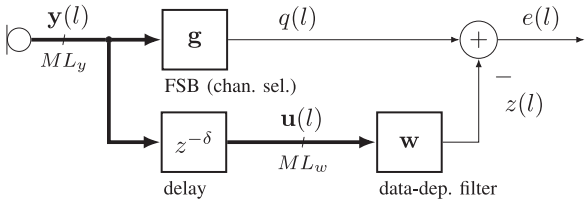


Fig. 1. The MCLP framework employing the prediction delay δ in the data-dependent filter path.

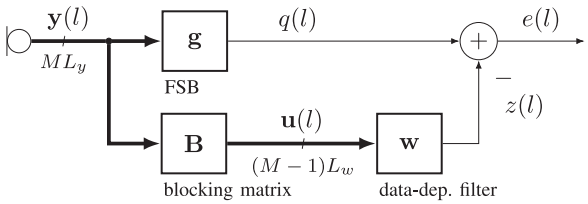


Fig. 2. The GSC framework employing the blocking matrix $\mathbf{B}$ in the data-dependent filter path.

therefore consists in estimating the multi-channel prediction filter from the microphone signals. When the prediction filter is of sufficient order, MCLP is theoretically able to completely equalize the RIRs [7]. Nowadays, MCLP is commonly implemented in frequency sub-bands using the short-time Fourier transform (STFT) [8]–[17], [26]–[28]. Incorporating the power spectral density (PSD) of the speech-source signal in the cost function has been shown to be beneficial, as, e.g., in the weighted prediction error (WPE) method [10], [11], where the speech-source signal is modeled as time-varying Gaussian [8]–[11] or using sparse priors [13]. Adaptive approaches based on recursive least squares [12], [15] and the Kalman filter [14], [16], [17], [28] have been proposed. In [17], given *noisy* microphone signals, the reverberant-speech component and the prediction-filter coefficients are estimated in an alternating fashion. To reduce noise after dereverberation, it has been proposed to cascade MCLP with minimum-variance distortionless response (MVDR) beamforming [26], [27], which became a popular approach in the recent CHiME-5 challenge [29].

Beamforming is designed to perform spatial filtering, and is hence commonly used for noise reduction, but may *also* be applied for dereverberation [30]. One can distinguish between data-independent (e.g., superdirective) beamforming and data-dependent (e.g., MVDR) beamforming. Although beamforming traditionally does *not* target channel inversion, it may be considered equivalent to MINT if the (presumably known) RIRs are incorporated in the filter design [18]. The so-called MINTFormer [20] provides a trade-off between the performance of MINT and the robustness of beamforming. In [25], a MINT-based multi-channel Wiener filter for joint dereverberation and noise reduction has been proposed. The analysis in [19] shows that for MVDR beamforming incorporating known RIRs, an inherent performance trade-off exists between dereverberation and noise reduction in case of incoherent as well as mixed coherent- and incoherent-noise fields. In this work, we are mainly concerned with the generalized sidelobe canceller (GSC) [31], [32], an implementation of the minimum-power distortionless response (MPDR) beamformer and widely employed in noise reduction. Fig. 2 depicts a block diagram of the GSC, which consists of three components: a filter-and-sum beamformer (FSB) steering a beam into the target direction, a blocking matrix blocking the speech component, and a data-dependent filter minimizing the output power and thereby suppressing residual noise components. The ability to block the speech component is essential to the GSC, as speech leakage through the blocking matrix may lead to partial speech cancellation. Employing the GSC for dereverberation, the blocking matrix should block the early (but not the late) reverberant-speech component. In [21] therefore a blocking matrix is used incorporating the relative early transfer functions (RETFs) of the speech source in order to jointly perform dereverberation and noise reduction. In the nested GSC [22], an inner GSC is employed for dereverberation and an outer GSC for noise reduction. In [28], we have proposed to integrate the GSC and MCLP in a parallel manner, and compared to the corresponding MCLP-GSC cascade, cf. also [26], [27].

A comparison of the block diagrams in Fig. 1 and 2 readily reveals the major difference between the two frameworks, which is due to their different objective. Where MCLP – designed for deconvolution – applies a simple delay to the microphone signals in the data-dependent filter path, the GSC instead – designed for spatial filtering – applies a blocking matrix. On the one hand, regarding dereverberation, the need for a blocking matrix is certainly a drawback of the GSC as compared to MCLP, as its design requires prior knowledge. On the other hand, regarding noise reduction, the blocking matrix distinguishes the speech source from potential localized noise sources, which is not possible in MCLP. For the noiseless dereverberation task, we have shown in [24] that the MCLP and GSC framework theoretically lead to the mathematically equivalent results for stationary source signals. In practice, additional noise may always be present. In this paper therefore, using pre-whitened least squares (LS) filter estimates, we formally analyze and compare the behavior of both frameworks in case of noise, both in terms of dereverberation and noise reduction. The main intention is to provide a better understanding of the theoretical performance limitations of both frameworks depending on a number of boundary conditions, such as noise levels, filter length and number of microphones, which cannot be done by naive comparison. In our theoretical analysis, we assume *complete blocking* of the early reverberant-speech component in the GSC blocking matrix, which requires prior knowledge of the early part of the speech-source RIRs or the RETFs. We derive that if the number of microphones is sufficiently large, the GSC theoretically achieves complete coherent-noise cancellation if incoherent noise is absent, while MCLP cancels the late coherent-noise components only, as expected by design. Further, in case of *complete blocking*, the GSC performs equally well as MCLP in terms of dereverberation; theoretically achieving complete reverberation cancellation if incoherent noise is absent. These theoretical findings are confirmed by time domain simulations. In addition, in case of *incomplete blocking*, based on STFT domain simulations using estimated RETFs, we show that the GSC instead performs inferior to MCLP in terms of dereverberation.

In Section II, the signal model for both frameworks is presented. In Section III, the filter estimation is discussed. Section IV and V proceed with the performance analysis of

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

the MCLP and the GSC framework, respectively. A comparative summary of the two frameworks is presented in Section VI, followed by simulation results in Section VII.

II. Signal Model

In this section, we define the signal model for both MCLP and the GSC. For simplicity, we employ the same notation for those signals and filters that correspond in both frameworks, cf. Fig. 1 and 2. As outlined before, the major difference between both consists in the use of a prediction delay δ in MCLP (cf. Fig. 1) and a blocking matrix $\mathbf{B}$ in the GSC (cf. Fig. 2) in the data-dependent filter path. In addition, the GSC speech reference is typically created by applying an FSB, whereas in MCLP a particular microphone signal is traditionally selected [6]–[16], [26], [27]. Both cases are covered generically by the filter $\mathbf{g}$ (cf. Fig. 1 and 2). The signal model equivalently applies in the time domain and the STFT domain, where l respectively denotes the time or frame index. In case of the STFT domain, throughout the paper, the frequency sub-band index is omitted as we treat all frequency sub-bands independently. Subsequently, vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{I}^{L\times L}$ and $\mathbf{0}^{L_1\times L_2}$ denote identity and zero matrices with the (optional) superscript indicating their dimensions, $\mathbf{A}^*$, $\mathbf{A}^T$, $\mathbf{A}^H$, $\mathbf{A}^+$, and $\mathrm{E}[\mathbf{A}]$ denote the complex conjugate, the transpose, the complex conjugate transpose, the pseudoinverse and the expected value of a matrix $\mathbf{A}$, $\mathrm{blkdiag}[\mathbf{A}_1,\ \ldots,\ \mathbf{A}_N]$ constructs a block-diagonal matrix from its arguments, and $\mathrm{tplz}[\mathbf{a},\ L]$ creates a Toeplitz matrix of L columns with the first column defined by the vector $(\mathbf{a}^T\ \mathbf{0}^{1\times(L-1)})^T$.

The acoustic scenario is presented in Section II-A, while in Section II-B the speech reference signal and its individual components are defined. In Section II-C and Section II-D, the data-dependent filter input signal is discussed for MCLP and the GSC, respectively. In Section II-E, the filter output and the enhanced signal are generically defined.

A. Acoustic Scenario

We assume an acoustic scenario comprising one speech source emitting the signal $s_1(l)$, and $N-1$ localized noise sources emitting the signals $s_n(l)$, $n=2\ldots N$, in a reverberant environment with M microphones. The mth microphone signal $y_m(l)$, $m=1\ldots M$, consists of the reverberant-speech component, reverberant-noise components, referred to as coherent-noise components hereafter, as well as an incoherent-noise component (originating from spatially uncorrelated noise, e.g., sensor noise), i.e.,

$$y_m(l)=\sum_{n=1}^{N}\underbrace{\sum_{k=0}^{L_h-1}h_{n,m}^*(k)s_n(l-k)}_{x_{n,m}(l)}+v_m(l), \tag{1}$$

with $h_{n,m}(k)$ denoting the time-invariant (sub-band) RIR between the nth source and the mth microphone of length L_h (neglecting the dead time common to all RIRs), k the tap index, $x_{n,m}(l)$ the reverberant components (reverberant-speech and coherent-noise components), and $v_m(l)$ the incoherent-noise component. Note that in the STFT case, the sub-band convolution model in (1) poses an approximation of the time-domain convolution [33], where the sub-band RIR length L_h is roughly R_{STFT} times smaller than the corresponding time domain RIR length, with R_{STFT} denoting the hop size in the STFT analysis [33]. We define the stacked multi-microphone vector $\mathbf{y}(l)\in\mathbb{C}^{ML_y}$,

$$\mathbf{y}(l)=\left(\mathbf{y}_1^T(l)\cdots\ \mathbf{y}_M^T(l)\right)^T, \tag{2}$$

$$\mathbf{y}_m(l)=\left(y_m(l)\cdots\ y_m(l-L_y\quad+\quad 1)\right)^T, \tag{3}$$

with L_y the number of samples/frames per microphone. With $\mathbf{x}_n(l)$ and $\mathbf{v}(l)$ defined in a similar manner as in (2), we obtain

$$\mathbf{y}(l)=\sum_{n=1}^{N}\mathbf{x}_n(l)+\mathbf{v}(l)=\mathbf{x}(l)+\mathbf{v}(l). \tag{4}$$

With the blockwise Toeplitz matrix $\mathbf{H}\in\mathbb{C}^{NL_s\times ML_y}$ and the stacked source-signal vector $\mathbf{s}(l)\in\mathbb{C}^{NL_s}$ defined by

$$\mathbf{H}=\begin{pmatrix}\mathbf{H}_{1,1}&\cdots&\mathbf{H}_{1,M}\\ \vdots&&\vdots\\ \mathbf{H}_{N,1}&\cdots&\mathbf{H}_{N,M}\end{pmatrix}=\begin{pmatrix}\mathbf{H}_1\\ \vdots\\ \mathbf{H}_N\end{pmatrix}, \tag{5}$$

$$\mathbf{H}_{n,m}=\mathrm{tplz}\Big[\left(h_{n,m}(0)\cdots\ h_{n,m}(L_h-1)\right)^T,\ L_y\Big], \tag{6}$$

$$\mathbf{s}(l)=\left(\mathbf{s}_1^T(l)\cdots\ \mathbf{s}_N^T(l)\right)^T, \tag{7}$$

$$\mathbf{s}_n(l)=\left(s_n(l)\cdots\ s_n(l-L_s\quad+\quad 1)\right)^T, \tag{8}$$

$$L_s=L_h+L_y-1, \tag{9}$$

the vector $\mathbf{x}(l)$ can then be written as

$$\mathbf{x}(l)=\sum_{n=1}^{N}\mathbf{H}_n^H\mathbf{s}_n(l)=\mathbf{H}^H\mathbf{s}(l). \tag{10}$$

We assume $\mathbf{s}_n(l)$ and $\mathbf{v}(l)$ to be mutually independent, i.e., with the correlation matrices $\mathbf{\Psi}_{s_n}(l)=\mathrm{E}[\mathbf{s}_n(l)\mathbf{s}_n^H(l)]$ and $\mathbf{\Psi}_v(l)$ equivalently, using (4), (7), (10), we find

$$\mathbf{\Psi}_s(l)=\mathrm{blkdiag}[\mathbf{\Psi}_{s_1}(l),\ \ldots,\ \mathbf{\Psi}_{s_N}(l)], \tag{11}$$

$$\mathbf{\Psi}_x(l)=\mathbf{H}^H\mathbf{\Psi}_s(l)\mathbf{H}, \tag{12}$$

$$\mathbf{\Psi}_y(l)=\mathbf{\Psi}_x(l)+\mathbf{\Psi}_v(l). \tag{13}$$

The matrices $\mathbf{\Psi}_{s_n}(l)$ are assumed to be invertible, such that $\mathbf{\Psi}_s^{-1}(l)=\mathrm{blkdiag}[\mathbf{\Psi}_{s_1}^{-1}(l),\ \ldots,\ \mathbf{\Psi}_{s_N}^{-1}(l)]$. Note that in the STFT domain, it is commonly assumed that $\mathrm{E}[s_1(l)s_1^*(l-k)]=0$ for $k\neq 0$ if the STFT hop size is sufficiently large, i.e., $\mathbf{\Psi}_{s_1}(l)$ becomes a diagonal matrix. Similar assumptions could be made for other source signals, but are not required in our analysis.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

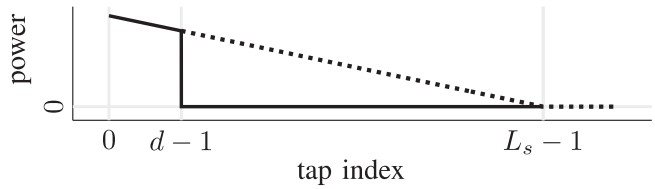


Fig. 3. Schematic of the (sub-band) IR relating $s_n(l)$ and $q_{s_n}(l)$, separated in early part $\dot{\mathbf{C}}\mathbf{H}_n\mathbf{g}$ [——] applied to $\mathbf{s}_n(l)$ and late part $\ddot{\mathbf{C}}_d\mathbf{H}_n\mathbf{g}$ [·····] applied to $\mathbf{s}_n(ld)$.

B. Speech Reference Signal

With the filter $\mathbf{g} \in \mathbb{C}^{ML_y}$, we define the speech reference signal $q(l)$ for both frameworks as

$$\begin{aligned} q(l) &= \mathbf{g}^H\mathbf{y}(l) \\ &= \sum_{n=1}^{N} \underbrace{(\mathbf{H}_n\mathbf{g})^H\mathbf{s}_n(l)}_{q_{s_n}(l)} + \underbrace{\mathbf{g}^H\mathbf{v}(l)}_{q_v(l)}, \end{aligned} \tag{14}$$

where $q_{s_n}(l)$ and $q_v(l)$ denote the individual source components of $q(l)$. Defining the parameter $d \ll L_h$ as the boundary between early and late reverberation, the reverberant-speech and coherent-noise components $q_{s_n}(l)$ may further be decomposed into early and late components $\dot{q}_{s_n}(l)$ and $\ddot{q}_{s_n}(l)$, i.e.,

$$q_{s_n}(l) = \underbrace{(\dot{\mathbf{C}}\mathbf{H}_n\mathbf{g})^H\mathbf{s}_n(l)}_{\dot{q}_{s_n}(l)} + \underbrace{(\ddot{\mathbf{C}}\mathbf{H}_n\mathbf{g})^H\mathbf{s}_n(l)}_{\ddot{q}_{s_n}(l)}, \tag{15}$$

with $\dot{\mathbf{C}} \in \mathbb{C}^{L_s \times L_s}$ and its complement $\ddot{\mathbf{C}}$ defined as

$$\dot{\mathbf{C}} = \begin{pmatrix} \mathbf{I}^{d\times d} & \mathbf{0}^{d\times(L_s-d)} \\ \mathbf{0}^{(L_s-d)\times d} & \mathbf{0}^{(L_s-d)\times(L_s-d)} \end{pmatrix}, \tag{16}$$

$$\ddot{\mathbf{C}} = \mathbf{I}^{L_s\times L_s} - \dot{\mathbf{C}}. \tag{17}$$

For later derivations throughout Section IV and Section V, we note that $\ddot{q}_{s_n}(l)$ in (15) may alternatively be expressed as

$$\ddot{q}_{s_n}(l) = (\ddot{\mathbf{C}}_d\mathbf{H}_n\mathbf{g})^H\mathbf{s}_n(l-d), \tag{18}$$

where $\ddot{\mathbf{C}}_d$ is derived from $\ddot{\mathbf{C}}$ by shifting d rows upwards, i.e.,

$$\ddot{\mathbf{C}}_d = \begin{pmatrix} \mathbf{0}^{(L_s-d)\times d} & \mathbf{I}^{(L_s-d)\times(L_s-d)} \\ \mathbf{0}^{d\times d} & \mathbf{0}^{d\times(L_s-d)} \end{pmatrix}. \tag{19}$$

Based on the above definitions, the (sub-band) impulse response (IR) relating $s_n(l)$ and $q_{s_n}(l)$ is graphically represented in Fig. 3. The parameter d can be controlled by design choices in the MCLP and the GSC framework, as shown in Section II-C and Section II-D, respectively.

We now define the early reverberant-speech component $\dot{q}_{s_1}(l)$ as the target component to be maintained, and the remaining late reverberant-speech component plus all noise components $\ddot{q}_{s_1}(l) + \sum_{n=2}^{N} q_{s_n}(l) + q_v(l)$ as the component to be canceled. Note that in a *different* acoustic scenario, e.g., with N speech sources instead of one speech source plus $N-1$ noise sources, the target component could be defined differently, e.g., by $\sum_{n=1}^{N} \dot{q}_{s_n}(l)$.

C. MCLP Filter Input

In the MCLP framework, the filter input signal $\mathbf{u}(l) \in \mathbb{C}^{ML_w}$ is a delayed version of the microphone signals $\mathbf{y}(l)$. The prediction delay δ is chosen as $\delta = d$, i.e.,

$$\boxed{\begin{aligned} \mathbf{u}(l) &= \mathbf{y}(l-d) \\ &= \mathbf{H}^H\mathbf{s}(l-d) + \mathbf{v}(l-d). \end{aligned}} \tag{20}$$

Hence, the length L_y in (9) equals the length L_w of a single filter channel of the data-dependent filter $\mathbf{w}$, i.e.,

$$L_y = L_w. \tag{21}$$

With (9), (21), we determine that $\mathbf{H} \in \mathbb{C}^{NL_s\times ML_y}$ is a fat matrix if the MCLP filter length L_w satisfies the condition

$$\boxed{L_w \geq \frac{N(L_h-1)}{M-N},} \tag{22}$$

which obviously requires $M > N$ microphones. If L_w is chosen according to (22) and the (sub-band) RIRs meet the MINT requirements (i.e., no common zeros), which is commonly assumed [2], [7], then the system is invertible and $\mathbf{H}$ has full row rank [2]. As it is crucial for our derivations in Section IV-B, full row rank of $\mathbf{H}$ is assumed in the remainder. Since our simulation results in Section VII support our theoretical conclusions in Section IV, we consider this assumption to be reasonable.

D. GSC Filter Input

In the GSC framework, the filter input signal $\mathbf{u}(l) \in \mathbb{C}^{(M-1)L_w}$ is constructed by applying a blocking matrix $\mathbf{B} \in \mathbb{C}^{ML_y\times(M-1)L_w}$ to the microphone signal, i.e.,

$$\boxed{\begin{aligned} \mathbf{u}(l) &= \mathbf{B}^H\mathbf{y}(l) \\ &= (\mathbf{H}\mathbf{B})^H\mathbf{s}(l) + \mathbf{B}^H\mathbf{v}(l), \end{aligned}} \tag{23}$$

where L_w again describes the length of a single filter channel of the filter $\mathbf{w}$. Eq. (23) is the GSC counterpart to (20) for MCLP. We intend to completely block all components in $\mathbf{x}_1(l) = \mathbf{H}_1^H\mathbf{s}_1(l)$ correlated to the target component $\dot{q}_{s_1}(l)$ as defined in (15). A matrix $\mathbf{B}$ satisfying[1] this condition may be defined in the following manner,

$$\mathbf{B} = \begin{pmatrix} -\dot{\mathbf{H}}_{1,2} \cdots -\dot{\mathbf{H}}_{1,M} \\ \text{blkdiag}\left[\dot{\mathbf{H}}_{1,1}, \ldots, \quad \dot{\mathbf{H}}_{1,1}\right] \end{pmatrix}, \tag{24}$$

$$\dot{\mathbf{H}}_{1,m} = \text{tplz}\Big[\big(h_{1,m}(0) \cdots h_{1,m}(L_b-1)\big)^T,\ L_w\Big], \tag{25}$$

where L_b denotes the length of the blocking filters, such that in the GSC, we find for L_y in (9),

$$L_y = L_b + L_w - 1. \tag{26}$$

[1]Many definitions of $\mathbf{B}$ achieving complete blocking exist. In the STFT domain, for $L_b = 1$, the blocking matrix may also be defined using RETFs [21]. If the target component is instead defined as $\sum_{n=1}^{N} \dot{q}_{s_n}(l)$ as in the *different* acoustic scenario mentioned in Section II-B, then also the definition of $\mathbf{B}$ needs to change accordingly.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

The definition in (24)–(25) ensures that all components corresponding to the first $d \geq L_b$ taps of the speech-source (sub-band) RIRs $h_{1,m}(k)$ are nullified, where the case $d > L_b$ occurs if the first L_b taps of $h_{1,m}(k)$ are succeeded by one or more zeros. The product $\mathbf{HB}$ takes the form

$$\mathbf{HB} = \begin{pmatrix} \mathbf{0}^{d\times(M-1)L_w} \\ \mathbf{H}_B \end{pmatrix}, \tag{27}$$

where, using (9), (26), $\mathbf{H}_B \in \mathbb{C}^{NL_s - d\times(M-1)L_w}$ is a fat matrix if the GSC filter length L_w satisfies the condition

$$\boxed{L_w \geq \frac{N(L_h-2)+(N-1)d}{M-N-1},} \tag{28}$$

which obviously requires $M > N+1$ microphones. If L_w is chosen according to (28) and the $M-1$ (sub-band) impulse responses in $\mathbf{HB}$ meet the MINT requirements, i.e., the nullity of $(\mathbf{HB})^H$ does not exceed d, then $\mathbf{H}_B$ has full row rank according to the rank-nullity theorem [34]. As it is crucial for our derivations in Section V-B1, full row rank of $\mathbf{H}_B$ is assumed in the remainder. Since our simulation results in Section VII support our theoretical conclusions in Section V, we consider this assumption to be reasonable. Comparing L_w for the GSC and MCLP in (28) and (22), respectively, we find that the GSC requires longer filters. Note however that the GSC employs one filter channel less.

E. Enhanced Signal

For both frameworks, the filter output signal $z(l)$ and the enhanced signal $e(l)$ are given by

$$z(l) = \mathbf{w}^H \mathbf{u}(l), \tag{29}$$

$$e(l) = q(l) - z(l), \tag{30}$$

with $\mathbf{u}(l)$ given by (20) in MCLP or (23) in the GSC. For MCLP, $z(l)$ is the linear prediction of $q(l)$, and $e(l)$ accordingly the linear prediction residual. The estimation of the filter $\mathbf{w}$ is discussed in Section III.

III. Filter Estimation

We now present the pre-whitened LS estimate of $\mathbf{w}$ in Section III-A and discuss the choice of the pre-whitening matrix in Section III-B. In Section III-C, we present the corresponding Wiener solution, which is then used in the theoretical analysis in the subsequent Section IV and Section V.

A. Pre-Whitened LS

With $l = 0 \ldots L_{obs} - 1$ and L_{obs} denoting the number of observations used in the filter estimation, let $\mathbf{q}_{(\cdot)} \in \mathbb{C}^{1\times L_{obs}}$ and $\mathbf{U}_{(\cdot)} \in \mathbb{C}^{ML_w \times L_{obs}}$ denote correspondingly stacked versions of $q(l)$ and $\mathbf{u}(l)$, i.e.,

$$\mathbf{q}_{(\cdot)} = \big(q(0) \cdots q(L_{obs}-1)\big), \tag{31}$$

$$\mathbf{U}_{(\cdot)} = \big(\mathbf{u}(0) \cdots \mathbf{u}(L_{obs}-1)\big), \tag{32}$$

and let $\mathbf{Y}_{(\cdot)}$, $\mathbf{S}_{(\cdot)}$, $\mathbf{S}_{n|(\cdot)}$, and $\mathbf{V}_{(\cdot)}$ be defined equivalently to $\mathbf{U}_{(\cdot)}$ in (32). Further, let $\mathbf{\Omega}_{(\cdot)}^{-1/2} \in \mathbb{C}^{L_{obs}\times L_{obs}}$ denote some pre-whitening matrix with $\mathbf{\Omega}_{(\cdot)} = \mathbf{\Omega}_{(\cdot)}^{H/2}\mathbf{\Omega}_{(\cdot)}^{1/2}$ to be defined explicitly in Section III-B. Let $\tilde{\mathbf{U}}_{(\cdot)}$ and $\tilde{\mathbf{q}}_{(\cdot)}$ denote correspondingly pre-whitened versions of $\mathbf{U}_{(\cdot)}$ and $\mathbf{q}_{(\cdot)}$, i.e.,

$$\tilde{\mathbf{q}}_{(\cdot)} = \mathbf{q}_{(\cdot)}\mathbf{\Omega}_{(\cdot)}^{-1/2} = \big(\tilde{q}(0) \cdots \tilde{q}(L_{obs}-1)\big)^T, \tag{33}$$

$$\tilde{\mathbf{U}}_{(\cdot)} = \mathbf{U}_{(\cdot)}\mathbf{\Omega}_{(\cdot)}^{-1/2} = \big(\tilde{\mathbf{u}}(0) \cdots \tilde{\mathbf{u}}(L_{obs}-1)\big), \tag{34}$$

and let $\tilde{\mathbf{Y}}_{(\cdot)}$, $\tilde{\mathbf{S}}_{(\cdot)}$, $\tilde{\mathbf{S}}_{n|(\cdot)}$, and $\tilde{\mathbf{V}}_{(\cdot)}$ as well as their respective column vectors $\tilde{\mathbf{y}}(l)$, $\tilde{\mathbf{s}}(l)$, $\tilde{\mathbf{s}}_n(l)$, and $\tilde{\mathbf{v}}(l)$ be defined equivalently to (34). Based on these definitions, the pre-whitened data $\tilde{q}(l)$ and $\tilde{\mathbf{u}}(l)$ may be expressed equivalently to $q(l)$ in (14) and $\mathbf{u}(l)$ in (20), (23), where $\tilde{\mathbf{y}}(l)$, $\tilde{\mathbf{s}}(l)$, and $\tilde{\mathbf{v}}(l)$ replace $\mathbf{y}(l)$, $\mathbf{s}(l)$, and $\mathbf{v}(l)$, respectively. Based on (29)–(30), (33)–(34), we generically define the LS cost function,

$$J_{LS}(\mathbf{w}) = \Big\| \underbrace{\tilde{\mathbf{q}}_{(\cdot)} - \mathbf{w}^H \tilde{\mathbf{U}}_{(\cdot)}}_{\tilde{\mathbf{e}}_{(\cdot)} = (\tilde{e}(0) \ \cdots \ \tilde{e}(L_{obs}-1))} \Big\|_2^2, \tag{35}$$

leading to the the LS filter estimate $\mathbf{w}_{LS}$,

$$\begin{aligned} \mathbf{w}_{LS} &= \arg\min_{\mathbf{w}} J_{LS}(\mathbf{w}) \\ &= \left(\tilde{\mathbf{U}}_{(\cdot)}\tilde{\mathbf{U}}_{(\cdot)}^H\right)^{-1} \tilde{\mathbf{U}}_{(\cdot)}\tilde{\mathbf{q}}_{(\cdot)}^H \\ &= \left(\mathbf{U}_{(\cdot)}\mathbf{\Omega}_{(\cdot)}^{-1}\mathbf{U}_{(\cdot)}^H\right)^{-1} \mathbf{U}_{(\cdot)}\mathbf{\Omega}_{(\cdot)}^{-1}\mathbf{q}_{(\cdot)}^H. \end{aligned} \tag{36}$$

Note that with (33)–(34), $\mathbf{w}_{LS}$ in (36) may alternatively be written as

$$\mathbf{w}_{LS} = \left(\sum_{l=0}^{L_{obs}-1} \tilde{\mathbf{u}}(l)\tilde{\mathbf{u}}^H(l)\right)^{-1} \sum_{l=0}^{L_{obs}-1} \tilde{\mathbf{u}}(l)\tilde{q}^*(l). \tag{37}$$

B. Choice of Pre-Whitening Matrix

The pre-whitening matrix $\mathbf{\Omega}_{(\cdot)}^{-1/2}$ may be used to mitigate an estimation bias due to the speech source-signal statistics. E.g., for the two cases $\boldsymbol{\varepsilon}_{(\cdot)} \in \{\dot{\mathbf{q}}_{s_1|(\cdot)}, \mathbf{s}_{1|(\cdot)}\}$ with $\dot{\mathbf{q}}_{s_1|(\cdot)}$ and $\mathbf{s}_{1|(\cdot)}$ defined equivalently to (31), an unbiased estimate is achieved if

$$\mathbf{\Omega}_{(\cdot)} = \mathbf{\Psi}_{\varepsilon|(\cdot)}, \tag{38}$$

where $\mathbf{\Psi}_{\varepsilon|(\cdot)} = \mathrm{E}[\boldsymbol{\varepsilon}_{(\cdot)}^H \boldsymbol{\varepsilon}_{(\cdot)}]$. Eq. (36) then corresponds to the (unbiased) generalized LS estimator of $\mathbf{w}$ for the data model $\mathbf{q}_{(\cdot)} = \mathbf{w}^H\mathbf{U}_{(\cdot)} + \boldsymbol{\varepsilon}_{(\cdot)}$, where $\boldsymbol{\varepsilon}_{(\cdot)}$ resembles the observation noise. In the time domain, $\mathbf{\Psi}_{\dot{s}_1|(\cdot)}$ and $\mathbf{\Psi}_{s_1|(\cdot)}$ are generally non-diagonal. The choice $\mathbf{\Omega}_{(\cdot)} = \mathbf{\Psi}_{s_1|(\cdot)}$ here corresponds to the pre-whitening paradigms proposed in [6] for MCLP and [23] for the GSC. The choice $\mathbf{\Omega}_{(\cdot)} = \mathbf{I}$ generally yields a biased estimate, as demonstrated in the simulations in Section VII-B1.

In the STFT domain, $\mathbf{\Psi}_{\dot{s}_1|(\cdot)}$ may be modeled as a matrix with $2d-1$ non-zero diagonals and $\mathbf{\Psi}_{s_1|(\cdot)}$ may be modeled as a fully diagonal matrix, cf. Section II-A, where the l^{th} diagonal element of $\mathbf{\Psi}_{s_1|(\cdot)}$ corresponds to the PSD $\psi_{s_1}(l) = \mathrm{E}[|s_1(l)|^2]$. In this case, with (31)–(32), $\mathbf{w}_{LS}$ in (36) may therefore be

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

written as

$$\mathbf{w}_{LS} = \left(\sum_{l=0}^{L_{obs}-1} \frac{\mathbf{u}(l)\mathbf{u}^H(l)}{\psi_{s_1}(l)} \right)^{-1} \sum_{l=0}^{L_{obs}-1} \frac{\mathbf{u}(l)q^*(l)}{\psi_{s_1}(l)}, \quad (39)$$

i.e., each frame $q(l)$ and $\mathbf{u}(l)$ is weighted by the inverse of $\psi_{s_1}(l)$, which, in case of MCLP, corresponds to the WPE criterion [10], [11]. Note that $\psi_{s_1}(l)$ varies over time for non-stationary source signals.

In Section VII-B2, we present STFT-domain simulations for $d = 1$ and $\mathbf{\Omega}_{(\cdot\cdot)} = \mathbf{\Psi}_{\dot{q}_{s_1}|(\cdot\cdot)} \propto \mathbf{\Psi}_{s_1|(\cdot\cdot)}$. Herein, prior to estimating $\mathbf{w}$ according to (36), the PSDs $\psi_{\dot{q}_{s_1}}(l)$ on the diagonal of $\mathbf{\Psi}_{\dot{q}_{s_1}|(\cdot\cdot)}$ are estimated as proposed in [28], [35], i.e., by applying the generalized eigenvalue decomposition (GEVD) to the spatial correlation matrix of the microphone signals in each frame l, where it is assumed that the spatial coherence matrix of the late reverberant component may be modeled as diffuse, cf. Section VII-A3.

Note that in the *different* acoustic scenario mentioned in Section II-B with the target component defined as $\sum_{n=1}^{N} \dot{q}_{s_n}(l)$ instead of $\dot{q}_{s_1}(l)$, in order to achieve an unbiased filter estimate, one has to change $\mathbf{\Omega}_{(\cdot\cdot)}$ accordingly, e.g., using $\boldsymbol{\varepsilon}_{(\cdot\cdot)} = \sum_{n=1}^{N} \dot{\mathbf{q}}_{s_n|(\cdot\cdot)}$ in (38).

C. Convergence to Wiener Filter Solution

For the purpose of the analysis in Section IV and Section V, we assume wide-sense stationarity for the pre-whitened signals $\tilde{\mathbf{u}}(l)$ and $\tilde{q}(l)$, i.e., their statistics are independent of l. Then, for $L_{obs} \to \infty$, the estimate $\mathbf{w}_{LS}$ in (37) converges to the Wiener filter solution $\mathbf{w}_{WF}$,

$$\mathbf{w}_{WF} = \mathbf{\Psi}_{\tilde{u}}^{+} \boldsymbol{\psi}_{\tilde{u}\tilde{q}}, \quad (40)$$

with $\mathbf{\Psi}_{\tilde{u}} = \mathrm{E}[\tilde{\mathbf{u}}(l)\tilde{\mathbf{u}}^H(l)]$ and $\boldsymbol{\psi}_{\tilde{u}\tilde{q}} = \mathrm{E}[\tilde{\mathbf{u}}(l)\tilde{q}^*(l)]$. Here, the inverse in (37) is replaced by the pseudoinverse, as in the GSC, $\mathbf{\Psi}_{\tilde{u}}$ becomes rank-deficient in absence of incoherent noise and in case of complete blocking, i.e., if (27) holds, cf. Section V-B1 and Appendix A-1.

IV. MCLP Analysis

For $\mathbf{w} = \mathbf{w}_{WF}$, we now derive the MCLP filter output signal $z(l)$ in Section IV-A and then derive and discuss the enhanced signal $e(l)$ under different noise conditions in Section IV-B.

A. MCLP Filter Output

Using (14), (20), and noting that $\mathrm{E}[\tilde{\mathbf{y}}(l-d)\tilde{\mathbf{y}}^H(l-d)] = \mathrm{E}[\tilde{\mathbf{y}}(l)\tilde{\mathbf{y}}^H(l)]$, the terms $\mathbf{\Psi}_{\tilde{u}}$ and $\boldsymbol{\psi}_{\tilde{u}\tilde{q}}$ in (40) become

$$\mathbf{\Psi}_{\tilde{u}} = \mathbf{\Psi}_{\tilde{y}}, \quad (41)$$

$$\boldsymbol{\psi}_{\tilde{u}\tilde{q}} = \mathbf{\Psi}_{\tilde{y}|d}\mathbf{g}, \quad (42)$$

$$\mathbf{\Psi}_{\tilde{y}|d} = \mathrm{E}\left[\tilde{\mathbf{y}}(l-d)\tilde{\mathbf{y}}^H(l)\right]. \quad (43)$$

Inserting (20) in (29) and substituting $\mathbf{w}$ by $\mathbf{w}_{WF}$ in (40), we obtain for the filter output signal,

$$z(l) = (\mathbf{\Psi}_{\tilde{y}}^{+} \mathbf{\Psi}_{\tilde{y}|d}\mathbf{g})^H \mathbf{y}(l-d). \quad (44)$$



(a) correlation matrix $\mathbf{\Psi}_{\tilde{s}_n}$ (b) correlation matrix $\mathbf{\Psi}_{\tilde{s}_n|d}$

Fig. 4. Schematic of the correlation matrices $\mathbf{\Psi}_{\tilde{s}_n}$ and $\mathbf{\Psi}_{\tilde{s}_n|d}$ as different submatrices of a larger correlation matrix.

Let the shifted correlation matrices $\mathbf{\Psi}_{\tilde{s}_n|d}$, $\mathbf{\Psi}_{\tilde{s}|d}$, and $\mathbf{\Psi}_{\tilde{v}|d}$ be defined equivalently to $\mathbf{\Psi}_{\tilde{y}|d}$ in (43), with relations equivalent to (11)–(13). We now introduce the following relation between $\mathbf{\Psi}_{\tilde{s}_n|d}$ and $\mathbf{\Psi}_{\tilde{s}_n}$, which is used in the subsequent derivations in Section IV-B. For this, note that we can interpret $\mathbf{\Psi}_{\tilde{s}_n|d}$ and $\mathbf{\Psi}_{\tilde{s}_n}$ as different submatrices of a larger correlation matrix, as shown in Fig. 4. The submatrix defining $\mathbf{\Psi}_{\tilde{s}_n|d}$ is shifted left by d columns as compared to the submatrix defining $\mathbf{\Psi}_{\tilde{s}_n}$. Noting that the autocorrelation width of $\tilde{s}_n(l)$ is typically much smaller than L_h in both the time and STFT domain, we assume that the autocorrelation of $\tilde{s}_n(l)$ is zero for lags greater than $L_s - d$, where $L_s - d \geq \frac{M}{M-N}(L_h - 1) - d$ and $d \ll L_h$, cf. (9), (21)–(22). Using (16), (19), we then express $\mathbf{\Psi}_{\tilde{s}_n|d}$ in terms of $\mathbf{\Psi}_{\tilde{s}_n}$ by

$$\mathbf{\Psi}_{\tilde{s}_n|d} = \mathbf{\Psi}_{\tilde{s}_n}\ddot{\mathbf{C}}_d + \ddot{\mathbf{C}}_d\mathbf{\Psi}_{\tilde{s}_n}\dot{\mathbf{C}}. \quad (45)$$

The product $\mathbf{\Psi}_{\tilde{s}_n}\ddot{\mathbf{C}}_d$ shifts the elements in $\mathbf{\Psi}_{\tilde{s}_n}$ right by d columns. The product $\ddot{\mathbf{C}}_d\mathbf{\Psi}_{\tilde{s}_n}\mathbf{C}$ replaces the resulting zero columns by the first d columns of $\mathbf{\Psi}_{\tilde{s}_n}$ shifted up by d rows.

B. MCLP Enhancement

We now analyze the behavior of MCLP considering two scenarios: absence and presence of incoherent noise.

1) Absence of Incoherent Noise: The absence of incoherent noise corresponds to $\mathbf{v}(l) = \mathbf{0}$, i.e., $\mathbf{y}(l) = \mathbf{x}(l)$. In this case, using (10) and relations equivalent to (11)–(13), the individual terms in (44) are equal to $\mathbf{y}(l-d) = \mathbf{H}^H\mathbf{s}(l-d)$, $\mathbf{\Psi}_{\tilde{y}} = \mathbf{H}^H\mathbf{\Psi}_{\tilde{s}}\mathbf{H}$, and $\mathbf{\Psi}_{\tilde{y}|d} = \mathbf{H}^H\mathbf{\Psi}_{\tilde{s}|d}\mathbf{H}$. Inserting these in (44) and noting that $\mathbf{H}^{+} = \mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$ and hence $\mathbf{H}\mathbf{H}^{+} = \mathbf{I}$ since $\mathbf{H}$ is assumed to have full row rank yields

$$z(l) = (\mathbf{\Psi}_{\tilde{s}}^{+}\mathbf{\Psi}_{\tilde{s}|d}\mathbf{H}\mathbf{g})^H\mathbf{s}(l-d), \quad (46)$$

which, using (11), (45), (14)–(15), may be written as

$$z(l) = \sum_{n=1}^{N} \left(\ddot{q}_{s_n}(l) + \boldsymbol{\vartheta}_{n|d}^H \mathbf{s}_n(l-d) \right), \quad (47)$$

$$\text{with} \quad \boldsymbol{\vartheta}_{n|d} = \mathbf{\Psi}_{\tilde{s}_n}^{-1}\ddot{\mathbf{C}}_d\mathbf{\Psi}_{\tilde{s}_n}\dot{\mathbf{C}}\mathbf{H}_n\mathbf{g}. \quad (48)$$

As apparent from (47)–(48), all reverberant source components are treated *mutually independently* and *equally*. This holds as long as (22) is satisfied and $\mathbf{H}$ has full row rank. Inserting (47)

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

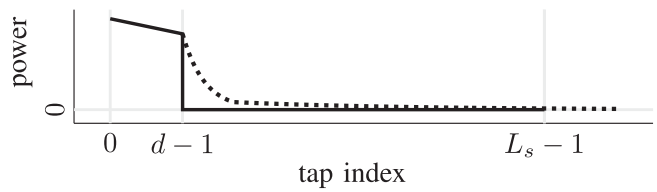


Fig. 5. Schematic of the (sub-band) IR relating $s_n(l)$ and $e_{s_n}(l)$, separated in early part $\dot{\mathbf{C}}\mathbf{H}_n\mathbf{g}$ [——] applied to $\mathbf{s}_n(l)$ and bias part $-\boldsymbol{\vartheta}_{n|d}$ [· · · · ·] applied to $\mathbf{s}_n(l-d)$.

into (30) yields the MCLP output signal,

$$\boxed{e(l) = \sum_{n=1}^{N} \left(\underbrace{\dot{q}_{s_n}(l) - \boldsymbol{\vartheta}_{n|d}^H \mathbf{s}_n(l-d)}_{e_{s_n}(l)} \right).} \tag{49}$$

From (49), we observe that $e(l)$ equals the sum of the early components $\dot{q}_{s_n}(l)$ and a (potential) bias term $-\boldsymbol{\vartheta}_{n|d}^H \mathbf{s}_n(l-d)$ per source, with $\boldsymbol{\vartheta}_{n|d} \in \mathbb{C}^{L_s}$ and $L_s = L_h + L_w - 1$ according to (9), (26). Therefore, as only the late components $\ddot{q}_{s_n}(l)$ are canceled, the MCLP framework suits best in the *different* acoustic scenario mentioned in Section II-B with the target component defined as $\sum_{n=1}^{N} \dot{q}_{s_n}(l)$ instead of $\dot{q}_{s_1}(l)$. Combining (15) and (18), we can compare the individual components $e_{s_n}(l)$ in (49) to

$$q_{s_n}(l) = \dot{q}_{s_n}(l) + (\ddot{\mathbf{C}}_d \mathbf{H}_n \mathbf{g})^H \mathbf{s}_n(l-d), \tag{50}$$

i.e., the bias term replaces the late component $\ddot{q}_{s_n}(l) = (\ddot{\mathbf{C}}_d \mathbf{H}_n \mathbf{g})^H \mathbf{s}_n(l-d)$. Similarly as for the (sub-band) IR $\mathbf{H}_n\mathbf{g}$ relating $s_n(l)$ and $q_{s_n}(l)$ in Fig. 3, we visualize the (sub-band) IR relating $s_n(l)$ and $e_{s_n}(l)$, composed of the early part $\dot{\mathbf{C}}\mathbf{H}_n\mathbf{g}$ and the bias part $-\boldsymbol{\vartheta}_{n|d}$, in Fig. 5. In the following we interpret the bias term in more detail, which has also partly been done in our previous work [24]. Firstly, from $\dot{\mathbf{C}}\mathbf{H}_n\mathbf{g}$ in (48), we observe that the bias term $-\boldsymbol{\vartheta}_{n|d}^H \mathbf{s}_n(l-d)$ depends on the first d taps of $\mathbf{H}_n\mathbf{g}$ only, i.e., on its early part, but *not* its late part. Secondly, we note that $\boldsymbol{\vartheta}_{n|d}$ depends on the correlation matrix $\boldsymbol{\Psi}_{\tilde{s}_n}$ of the pre-whitened version $\tilde{\mathbf{s}}_n(l)$ of $\mathbf{s}_n(l)$, cf. Section III-A. We can hence argue that for $\boldsymbol{\Omega}_{(\cdot\cdot)} = \boldsymbol{\Psi}_{\dot{q}_{s_1}|(\cdot\cdot)}$ as defined in (38), with $\dot{q}_{s_1}(l) = (\dot{\mathbf{C}}\mathbf{H}_1\mathbf{g})^H \mathbf{s}_1(l)$, cf. (15), the coloration of the pre-whitened speech-source signal $\tilde{\mathbf{s}}_1(l)$ is inverse to the filter $\dot{\mathbf{C}}\mathbf{H}_1\mathbf{g}$, such that only the first element of the vector $\boldsymbol{\Psi}_{\tilde{s}_1}\dot{\mathbf{C}}\mathbf{H}_1\mathbf{g}$ is non-zero. Similarly, we can argue that for $\boldsymbol{\Omega}_{(\cdot\cdot)} = \boldsymbol{\Psi}_{s_1|(\cdot\cdot)}$, the matrix $\boldsymbol{\Psi}_{\tilde{s}_1}$ becomes diagonal, such that only the first d elements of $\boldsymbol{\Psi}_{\tilde{s}_1}\dot{\mathbf{C}}\mathbf{H}_1\mathbf{g}$ are non-zero. In both cases, with $\ddot{\mathbf{C}}_d$ as in (19), we find $\ddot{\mathbf{C}}_d\boldsymbol{\Psi}_{\tilde{s}_1}\dot{\mathbf{C}}\mathbf{H}_1\mathbf{g} = \mathbf{0}$, and therefore $\boldsymbol{\vartheta}_{1|d} = \mathbf{0}$ in (48) and finally $e_{s_1}(l) = \dot{q}_{s_1}(l)$ in (49). Hence, the estimator is indeed unbiased for $\boldsymbol{\Omega}_{(\cdot\cdot)} \in \{\boldsymbol{\Psi}_{\dot{q}_{s_1}|(\cdot\cdot)}, \boldsymbol{\Psi}_{s_1|(\cdot\cdot)}\}$, as anticipated in Section III.

Note that the remaining early components may still be biased, i.e., $\boldsymbol{\vartheta}_{n|d} \neq \mathbf{0}$ for $n \neq 1$. In general, for $\boldsymbol{\Omega}_{(\cdot\cdot)} = \mathbf{I}$, the term $\boldsymbol{\vartheta}_{n|d}^H \mathbf{s}_n(l-d)$ in (49) represents a (delayed) linear prediction component of $\dot{q}_{s_n}(l)$, i.e., the output signal component $e_{s_n}(l)$ may be understood as a (partially) whitened version of $\dot{q}_{s_n}(l)$. This effect is also known as excessive whitening [7].

2) Presence of Incoherent Noise: If additional incoherent noise $\mathbf{v}(l) \neq \mathbf{0}$ is present, the pseudoinverse of the sum $\boldsymbol{\Psi}_{\tilde{y}} = \boldsymbol{\Psi}_{\tilde{x}} + \boldsymbol{\Psi}_{\tilde{v}}$ in the filter $\boldsymbol{\Psi}_{\tilde{y}}^{+}\boldsymbol{\Psi}_{\tilde{y}|d}\mathbf{g}$ in (44) cannot be decomposed into its individual components, such that further simplification of (44) is not possible. In this more general case, MCLP cancels the linear prediction of the sum of $\mathbf{g}^H\mathbf{x}(l)$ and $\mathbf{g}^H\mathbf{v}(l)$. Noting that the incoherent noise acts as M additional independent sources, we find that the condition (22) for complete linear prediction in the MCLP framework, where M is required to exceed the number of independent sources N, cannot be fulfilled, resulting in decreased performance.

V. GSC Analysis

Similarly to Section IV, for $\mathbf{w} = \mathbf{w}_{WF}$, we now derive the GSC filter output signal $z(l)$ in Section V-A and then derive and discuss the enhanced signal $e(l)$ under different noise conditions in Section V-B.

A. GSC Filter Output

Following a derivation similar to Section IV-A, using (14) and (23), $\boldsymbol{\Psi}_{\tilde{u}}$ and $\boldsymbol{\psi}_{\tilde{u}\tilde{q}}$ in (40) can be written as

$$\boldsymbol{\Psi}_{\tilde{u}} = \mathbf{B}^H \boldsymbol{\Psi}_{\tilde{y}} \mathbf{B}, \tag{51}$$

$$\boldsymbol{\psi}_{\tilde{u}\tilde{q}} = \mathbf{B}^H \boldsymbol{\Psi}_{\tilde{y}} \mathbf{g}. \tag{52}$$

Inserting (23) in (29) and substituting $\mathbf{w}$ by $\mathbf{w}_{WF}$ in (40), we obtain for the filter output signal,

$$z(l) = \left(\mathbf{B}(\mathbf{B}^H \boldsymbol{\Psi}_{\tilde{y}} \mathbf{B})^{+} \mathbf{B}^H \boldsymbol{\Psi}_{\tilde{y}} \mathbf{g}\right)^H \mathbf{y}(l). \tag{53}$$

B. GSC Enhancement

Similarly to Section IV-B, we now analyze the behavior of the GSC, again considering two scenarios: absence and presence of incoherent noise.

1) Absence of Incoherent Noise: For the case $\mathbf{v}(l) = \mathbf{0}$, i.e., $\mathbf{y}(l) = \mathbf{x}(l)$, using (10) and relations equivalent to (11)–(13), the individual terms in (53) are equal to $\mathbf{y}(l) = \mathbf{H}^H\mathbf{s}(l)$ and $\boldsymbol{\Psi}_{\tilde{y}} = \mathbf{H}^H\boldsymbol{\Psi}_{\tilde{s}}\mathbf{H}$. Inserting in (53) yields

$$z(l) = \left(\mathbf{H}\mathbf{B}\left((\mathbf{H}\mathbf{B})^H \boldsymbol{\Psi}_{\tilde{s}} \mathbf{H}\mathbf{B}\right)^{+} (\mathbf{H}\mathbf{B})^H \boldsymbol{\Psi}_{\tilde{s}} \mathbf{H}\mathbf{g}\right)^H \mathbf{s}(l). \tag{54}$$

In Appendix A-1, assuming complete blocking such that (27) holds, it is shown that (54) can be reformulated as

$$\boxed{z(l) = \ddot{q}_{s_1}(l) + \boldsymbol{\vartheta}_{1|d}^H \mathbf{s}_1(l-d) + \sum_{n=2}^{N} q_{s_n}(l),} \tag{55}$$

$$\boxed{\text{with} \quad \boldsymbol{\vartheta}_{1|d} = (\ddot{\mathbf{C}}_d \boldsymbol{\Psi}_{\tilde{s}_1} \ddot{\mathbf{C}}_d^H)^{+} \ddot{\mathbf{C}}_d \boldsymbol{\Psi}_{\tilde{s}_1} \dot{\mathbf{C}}\mathbf{H}_0\mathbf{g}.} \tag{56}$$

Inserting (55) into (30) yields the GSC output signal,

$$e(l) = \dot{q}_{s_1}(l) - \boldsymbol{\vartheta}_{1|d}^H \mathbf{s}_1(l-d). \tag{57}$$

Eqs. (55)–(57) form the GSC counterpart to (47)–(49) for MCLP. From (57), we observe that $e(l)$ consists of two terms: the target component $\dot{q}_{s_1}(l)$ and a bias term $-\boldsymbol{\vartheta}_{1|d}^H \mathbf{s}_1(l-d)$. This implies that not only the late reverberant-speech component, but

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARATIVE SUMMARY OF THE MCLP FRAMEWORK VERSUS THE GSC FRAMEWORK

Property	**MCLP framework**	**GSC framework**
spatial knowledge	not required	required in $\mathbf{B}$, cf. (24)–(25)
filter input signal	$\mathbf{u}(l) = \mathbf{y}(l-d)$	$\mathbf{u}(l) = \mathbf{B}^H\mathbf{y}(l)$
filter length required for complete* cancellation	$L_w \geq \frac{N(L_h-1)}{M-N}$, requires $M > N$	$L_w \geq \frac{N(L_h-2)+(N-1)d}{M-N-1}$, requires $M > N+1$
output signal*	$e(l) = \sum_{n=1}^{N} \dot{q}_{s_n}(l) - \boldsymbol{\vartheta}_{n\|d}^H \mathbf{s}_n(l-d)$	$e(l) = \dot{q}_{s_1}(l) - \boldsymbol{\vartheta}_{1\|d}^H \mathbf{s}_0(l-d)$

*if incoherent noise absent (reduced performance otherwise).

also the coherent-noise components are completely canceled in the GSC, which is in contrast to MCLP, where only the late, but not the early coherent-noise components could be canceled. In Appendix A-2, it is shown that $\boldsymbol{\vartheta}_{1|d}$ in (56) for the GSC is indeed equal to $\boldsymbol{\vartheta}_{1|d}$ in (48) for MCLP. Hence, the discussion on the bias term in MCLP in Section IV-B1 similarly applies to the GSC, implying that for $\boldsymbol{\Omega}_{(\cdot\cdot)} \in \{\boldsymbol{\Psi}_{\dot{q}_{s_1}|(\cdot\cdot)}, \boldsymbol{\Psi}_{s_1|(\cdot\cdot)}\}$ in (36), we find $\boldsymbol{\vartheta}_{1|d} = \mathbf{0}$ and $e(l) = \dot{q}_{s_1}(l)$, i.e., we achieve complete and unbiased cancellation. Note that if the target component was defined as $\sum_{n=1}^{N} \dot{q}_{s_n}(l)$ instead of $\dot{q}_{s_1}(l)$ as in the *different* acoustic scenario mentioned in Section II-B and $\mathbf{B}$ was changed accordingly, for the same $\boldsymbol{\Omega}_{(\cdot\cdot)}$, the GSC would yield the same result as MCLP. Also note that the above conclusions hold for complete blocking, i.e., if (27) is satisfied. For *incomplete blocking*, partial speech cancellation may appear. In Section VII, we simulate both cases.

2) Presence of Incoherent Noise: Similarly to the MCLP framework, if additional incoherent noise $\mathbf{v}(l) \neq \mathbf{0}$ is present, a simplification of (53) is not possible. We may therefore apply the same reasoning as in Section IV-B2. Noting that the incoherent noise acts as M additional independent noise sources, we find that the condition for complete cancellation in the GSC framework (28), where $M-1$ is required to exceed the number of independent sources N, cannot be fulfilled, resulting in decreased performance. These conclusions are compliant with the analysis in [19], which demonstrates that in MVDR beamforming, there is an inherent trade-off between dereverberation and noise reduction for incoherent and mixed-coherent-plus-incoherent noise fields.

VI. COMPARATIVE SUMMARY

Table I summarizes the theoretical findings from Section IV and Section V. For a given pre-whitening matrix $\boldsymbol{\Omega}_{(\cdot\cdot)}^{-1/2}$, MCLP does not further distinguish between the speech source and the localized noise sources, while the GSC does so by means of the spatial pre-processing in the blocking matrix. MCLP hence treats all source components $q_{s_n}(l)$ the same, suppressing the late components $\ddot{q}_{s_n}(l)$, but none of the early components $\dot{q}_{s_n}(l)$. By contrast, the GSC suppresses all coherent-noise components $q_{s_n}(l)$ on the one hand, and the late reverberant-speech component $\ddot{q}_{s_1}(l)$ on the other hand, provided that the blocking matrix $\mathbf{B}$ achieves *complete blocking* of the early reverberant-speech component. In both frameworks, an unbiased speech component with $\boldsymbol{\vartheta}_{1|d} = \mathbf{0}$ may be obtained by pre-whitening $q(l)$ and $\mathbf{u}(l)$ accordingly. The presence of incoherent noise decreases the performance.

The spatial pre-processing of the GSC naturally comes at a cost. The blocking matrix requires spatial information, which needs to be acquired in practice. Further, as to be demonstrated in simulations, cf. Section VII-B, in case of *incomplete blocking*, the GSC performs inferior in terms of dereverberation as compared to MCLP. Compared to MCLP, the minimum number of required microphones is increased by one, as the blocking matrix creates $M-1$ independent output signals only from M input signals. In the GSC, the number of filter channels is accordingly decreased by one, where a higher filter length L_w is required per channel.

VII. SIMULATIONS

In this section, we present simulation results comparing MCLP and the GSC in terms of dereverberation and noise reduction performance. The simulation setup is described in VII-A, and the results are discussed in VII-B.

A. Simulation Setup

In order to confirm the theory in Section IV–V and to further assess the practical relevance we respectively perform simulations for time domain and STFT domain implementations. The reasons for this are as follows: in the time domain, using oracle knowledge on the early RIRs, complete blocking can be simulated for the GSC, cf. Section VII-A3a. Further, unweighted global power-ratio measures can be well defined and evaluated, cf. Section VII-A4a. Therefore, in order to confirm the theory, we perform simulations on the time domain implementation, employing an ideal blocking matrix in the GSC yielding *complete blocking*, and evaluate the performance using unweighted global power-ratio measures. In the STFT domain, complete blocking cannot be simulated, since the sub-band convolution model in (1) poses an approximation of the time-domain convolution only [33]. Instead of using oracle knowledge in the blocking matrix, we estimate the RETFs from the microphone signals, such that the GSC performance also depends on the estimation quality of the RETFs, cf. Section VII-A3b. As, due to incomplete blocking, power-ratio measures equivalent to those used in the time domain cannot be well defined, and as unweighted

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

global power-ratio measures are further known to relate comparably poorly to the perceived speech quality, we instead use perceptually motivated freqeuncy-weighted segmental power-ratio measures [36]. Therefore, in order to address the practical relevance, we perform simulations on the STFT domain implementation, employing an estimated blocking matrix in the GSC yielding *incomplete blocking*, and evaluate the performance using weighted segmental power-ratio measures.

1) Acoustic Scenario: In order to generate multi-channel RIRs, the randomized image method [37] is used at a sampling frequency of 16 kHz, whereby the image sources are randomly displaced within a sphere of 8 cm. Multi-channel RIRs are generated using the randomized image method [37] at a sampling frequency of 16 kHz, with the image sources randomly displaced within a sphere of 8 cm. A fractional delay low-pass filter with a relative cut-off frequency of 0.9 and a length of 11 taps is applied, such that the energy of each acoustic wave, i.e., of the direct component and each reflection, is spread over 11 samples. The room dimensions are $5\times4\times3$ m, the reverberation time is 0.5 s. The room impulse responses are truncated after 8000 taps. A linear array of 8 microphones with inter-microphone distances of (4, 4, 4, 8, 4, 4, 4) cm is used. The simulations comprise one speech source and one localized noise source. In total, 128 scenarios are generated. In each scenario, the position and orientation of the microphone array is randomized. The speech source is located at a random position in broadside direction at 2 m distance to the center of the microphone array (i.e., on a circle around its axis). The position of the localized noise source is randomized, with the constraint that the distance to the center of the microphone array is at least 1 m and the angle between the localized noise source and the speech source, seen from the center of the microphone array, is at least 15°.

2) Source Signals: We define two different source signal settings. In the first one, both the speech-source signal and the localized noise source signal are chosen to be temporally correlated, i.e., *colored* signals. The (non-stationary) speech-source signal $s_1(l)$ is composed of male and female speech of in total 51 s duration [38], while for the localized noise source signal $s_2(l)$, stationary pink noise is used. This setting is evaluated for both the time domain and the STFT domain implementations. In the time domain, for the chosen setup, cf. Section VII-A3, the coloration of the source signals causes a biased filter estimate. In the second setting, both $s_1(l)$ and $s_2(l)$ are chosen to be temporally uncorrelated, i.e., *white*, stationary signals, and have been generated independently from the source signals in the first setting. This setting is evaluated in the time domain implementations only, leading to an unbiased filter estimate and serving as a reference in order to illustrate the effect of the bias in the first setting. Since sensor noise is always present in practice, spatially and temporally uncorrelated noise $\mathbf{v}(l)$ is added in all simulations. Note that due to the incoherent noise, the time domain simulation results may at most approximately reach the theoretical limits discussed in Section IV-B1 and Section V-B1.

The power of the noise components is defined via the signal-to-coherent-noise ratio SNR_y^{coh} and the signal-to-incoherent-noise ratio SNR_y^{inc} in the first microphone, i.e.,

$$SNR_y^{coh} = 10\log_{10}\frac{\sum_l |x_{1,1}(l)|^2}{\sum_l |x_{2,1}(l)|^2}\,\text{dB}, \tag{58}$$

$$SNR_y^{inc} = 10\log_{10}\frac{\sum_l |x_{1,1}(l)|^2}{\sum_l |v_1(l)|^2}\,\text{dB}, \tag{59}$$

where the reverberant-speech component $x_{1,1}(l)$ in the first microphone is considered to be the useful signal.

3) MCLP and GSC Implementation:

a) Time domain: In the time domain, we define the direct speech component to be the target component, i.e., we choose $\delta = L_b = 11$ samples, corresponding to the energy spread of a single acoustic wave, cf. Section VII-A1, yielding $d = 11$ for MCLP and $d \geq 11$ for the GSC, cf. Section II-C and Section II-D. An ideal GSC blocking matrix $\mathbf{B}$ was designed, cf. (24)–(25). The filter $\mathbf{g}$ is chosen to be a matched filter (MF) such that $\mathbf{B}^H\mathbf{g} = \mathbf{0}$, both for the GSC and MCLP. We choose $\mathbf{\Omega}_{(\cdot)} = \mathbf{I}$ in (36), leading to a biased filter estimate for colored source signals. The effect of the bias is shown by comparing the performance for both colored and white source signals.

b) STFT domain: In the STFT domain, using square-root-Hann windows of 512 samples with 50% overlap, we choose $\delta = L_b = 1$ frame. The GSC blocking matrix $\mathbf{B}$ uses an estimate of the RETFs, which we obtain as presented in [28], [35], [39]: we estimate the *average* spatial correlation matrix of the microphone signals using the whole batch, and the spatial correlation matrix of the stationary noise components using 5 s noise-only frames, such that the spatial speech-component correlation matrix can be estimated by subtraction. Then, from the spatial speech-component correlation matrix estimate, the RETF relative to the first microphone is estimated using the GEVD, assuming that the spatial coherence matrix of the late reverberant-speech component in frame l may be modeled as diffuse [28], [35]. Again, the filter $\mathbf{g}$ is chosen to be an MF with $\mathbf{B}^H\mathbf{g} = \mathbf{0}$, i.e., $\mathbf{g}$ is a normalized version of the RETF estimate. For $\mathbf{\Omega}_{(\cdot)}$ in (36), we use an estimate of $\mathbf{\Psi}_{\dot{q}_{s_1}|(\cdot)}$, which in the STFT domain is diagonal for $d = 1$, cf. Section III-B. Since $\mathbf{g}$ is a normalized version of the RETF estimate, estimating the PSDs $\psi_{\dot{q}_{s_1}}(l)$ in $\mathbf{\Psi}_{\dot{q}_{s_1}|(\cdot)}$ corresponds to estimating the early-reverberant speech component in the first microphone. Again, this can be done using the GEVD [28], [35], now applied to a *recursive* estimate of the spatial speech-component correlation matrix in frame l. See, e.g., [28], [35] for a detailed and more formal discussion on GEVD-based RETF and PSD estimation.

4) Performance Measures:

a) Time domain: In the time domain, equivalently to $q_{s_1}(l)$, $q_{s_2}(l)$, $q_v(l)$ and $\dot{q}_{s_1}(l)$, we define the individual source components of $e(n)$ as $e_{s_1}(l)$, $e_{s_2}(l)$, $e_v(l)$ and $\dot{e}_{s_1}(l)$, where $\dot{e}_{s_1}(l) = \dot{q}_{s_1}(l)$, cf. (49), (57). With $\sigma \in \{q, e\}$, the signal-to-coherent-noise ratio SNR_σ^{coh}, the signal-to-incoherent-noise ratio SNR_σ^{inc}, the signal-to-total-noise ratio SNR_σ^{tot}, and the signal-to-reverberation ratio SRR_σ at the MF output and the

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

MCLP and GSC output are defined as

$$SNR_{\sigma}^{tot} = 10\log_{10}\frac{\sum_l |\dot{\sigma}_{s_1}(l)|^2}{\sum_l |\sigma_{s_2}(l)+\sigma_v(l)|^2}\,\text{dB}, \tag{60}$$

$$SRR_{\sigma} = 10\log_{10}\frac{\sum_l |\dot{\sigma}_{s_1}(l)|^2}{\sum_l |\sigma_{s_1}(l)-\dot{\sigma}_{s_1}(l)|^2}\,\text{dB}, \tag{61}$$

where the component $\dot{\sigma}_{s_1}(l)$ is considered to be the useful signal. Please note that $q(l)$, and hence for $\sigma = q$ also the measures in (58)–(61), are independent of the particular framework. Further, note that in the denominator of (61), for $\sigma = q$, the difference $q_{s_1}(l) - \dot{q}_{s_1}(l)$ equals the late reverberant-speech component $\ddot{q}_{s_1}(l)$, while for $\sigma = e$, the difference $e_{s_1}(l) - \dot{e}_{s_1}(l)$ comprises not only residual reverberation, but also a bias term in the general case. For evaluation, we use the improvement in SNR^{tot} and SRR, i.e.,

$$\Delta SNR^{tot} = SNR_e^{tot} - SNR_q^{tot}, \tag{62}$$

$$\Delta SRR = SRR_e - SRR_q. \tag{63}$$

b) STFT domain: In the STFT domain, the target component $\dot{q}_{s_1}(l)$ cannot be observed separately, since the subband convolution model in (1) poses an approximation of the time-domain convolution only [33]. Further, due to overlapping frames in the STFT processing and incomplete blocking in the GSC, the target component $\dot{q}_{s_1}(l)$ may not be completely maintained in $e(l)$, such that the measures in (60)–(63) are not suitable in the STFT domain. Instead, we define the direct-component in $q(l)$ as a reference signal, which *cannot* be assumed to be equivalent to $\dot{q}_{s_1}(l)$. Then, with $\sigma \in \{q, e\}$, for $\sigma(l)$ and $\sigma_{s_1}(l)$, respectively, we compute the frequency-weighted segmental signal-to-noise-plus-reverberation ratio and the frequency-weighted segmental signal-to-reverberation ratio [36], denoted as $SNRR^{fwseg}$ and SRR^{fwseg} and indicating the dereverberation-plus-noise-reduction performance and the dereverberation-only performance.

5) Varied Parameters: In the time domain, simulations are carried out for different values of the following parameters: SNR_y^{coh}, SNR_y^{inc}, L_w, and M. The filter length L_w is presented relatively to the theoretical minimum given in (22), (28), denoted by L_w^{rel}. While one parameter is varied, the others are fixed at $SNR_q^{coh} = 0$ dB, $SNR_q^{inc} = 90$ dB, $L_w^{\text{rel}} = 1$, and $M = 8$, i.e., all simulations intersect at this point. For $N = 2$, the minimum number of microphones required by MCLP and the GSC is given by $M = 3$ and $M = 4$, respectively, cf. (22), (28). If the number of microphones M falls below this required minimum, the filter length is computed setting the denominators in (22), (28) to one. Simulations posing nearly ideal conditions, i.e., sufficiently high SNR_q^{inc}, $L_w^{\text{rel}} \geq 1$ and sufficiently many microphones M, validate the theoretical results in Section IV and Section V, with minor deviations occurring due to the LS approximation in (36) of the Wiener solution in (40) and remaining low-level incoherent noise.

In the STFT domain, simulations are carried out for different values of SNR_y^{coh} only, with $SNR_q^{inc} = 90$ dB, $L_w^{\text{rel}} = 1$, and $M = 8$. Since complete blocking is not achieved in the STFT domain, decreased performance is expected for the GSC.

B. Simulation Results

We now discuss the time and STFT domain simulation results in Section VII-B1 and Section VII-B2, respectively.

1) Time Domain: The performance of both frameworks in terms of ΔSRR and ΔSNR^{tot} are shown in Fig. 6. We first discuss the dereverberation performance, followed by the noise reduction performance.

a) Dereverberation: From Fig. 6(a)–(d) we observe that under favorable conditions with predominantly late-reverberant-speech interference, i.e., for high SNR_y^{coh}, high SNR_y^{inc}, $L_w^{\text{rel}} \geq 1$ and sufficiently high M, the SRR improvement of both MCLP and GSC converge to the same value of around 31 dB for white source signals, respectively denoted by [·····] and [·····]. This upper limit is determined by the LS approximation (36) of the Wiener Solution of the (40). In all conditions, for colored source signals, the target component $\dot{q}_{s_1}(l)$ is partially whitened due to the biased filter estimate, leading to a performance drop for both MCLP [——] and the GSC [——], cf. Section IV-B1 and Section V-B1 and Table I. The GSC reaches up to 18 dB ΔSRR, outperforming MCLP by 3 dB. This is due to the potential delay between the direct component and the first reflection, increasing d for the GSC and thereby decreasing the bias, cf. Section VII-A3. The higher standard deviation of ΔSRR for the GSC is a result of the variation of this delay over different source and microphone array positions.

As can be seen in Fig. 6(a), for white source signals, MCLP shows a rather high sensitivity towards coherent noise for $SNR_y^{coh} < 10$ dB, whereas the GSC is somewhat less sensitive. This can be explained by the limited number of observations L_{obs} used in the LS approximation (36) of the Wiener solution (40), causing LS to focus on the suppression of components with higher power, i.e., here on (late) coherent-noise suppression. For colored source signals, the effect is less pronounced in both frameworks.

While it can be observed from Fig. 6(b) that both MCLP and GSC are highly sensitive to incoherent noise, the results also indicate an up to 2.5 dB lower performance of the GSC as compared to MCLP for $SNR_y^{inc} < 90$ dB and < 55 dB for white and colored source signals, respectively. The reason for this may lie in the GSC blocking matrix, which by construction causes a cross-correlation of the incoherent-noise components in the data-dependent filter input $\mathbf{u}(l)$, as opposed to the mere delay in MCLP. Hence, for the GSC, not only the autocorrelation submatrices of $\mathbf{\Psi}_{\tilde{u}}$ are affected by incoherent noise, but also the cross-correlation submatrices.

As shown in Fig. 6(c), ΔSRR saturates for both MCLP and GSC above $L_w^{\text{rel}} = 1$, both for white and colored source signals, as expected from theory. The GSC performs slightly better than MCLP if $L_w^{\text{rel}} < 1$. We may however state that for both frameworks, undermodeling is not extremely critical, as even at $L_w^{\text{rel}} = 0.7$, we obtain ΔSRR values above 22 dB for white source signals, while the performance is hardly affected for colored source signals.

From Fig. 6(d), we note that ΔSRR drops sharply for both MCLP and GSC if the number of microphones is smaller than required, i.e., $M < 3$ and $M < 4$, respectively. This holds for

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

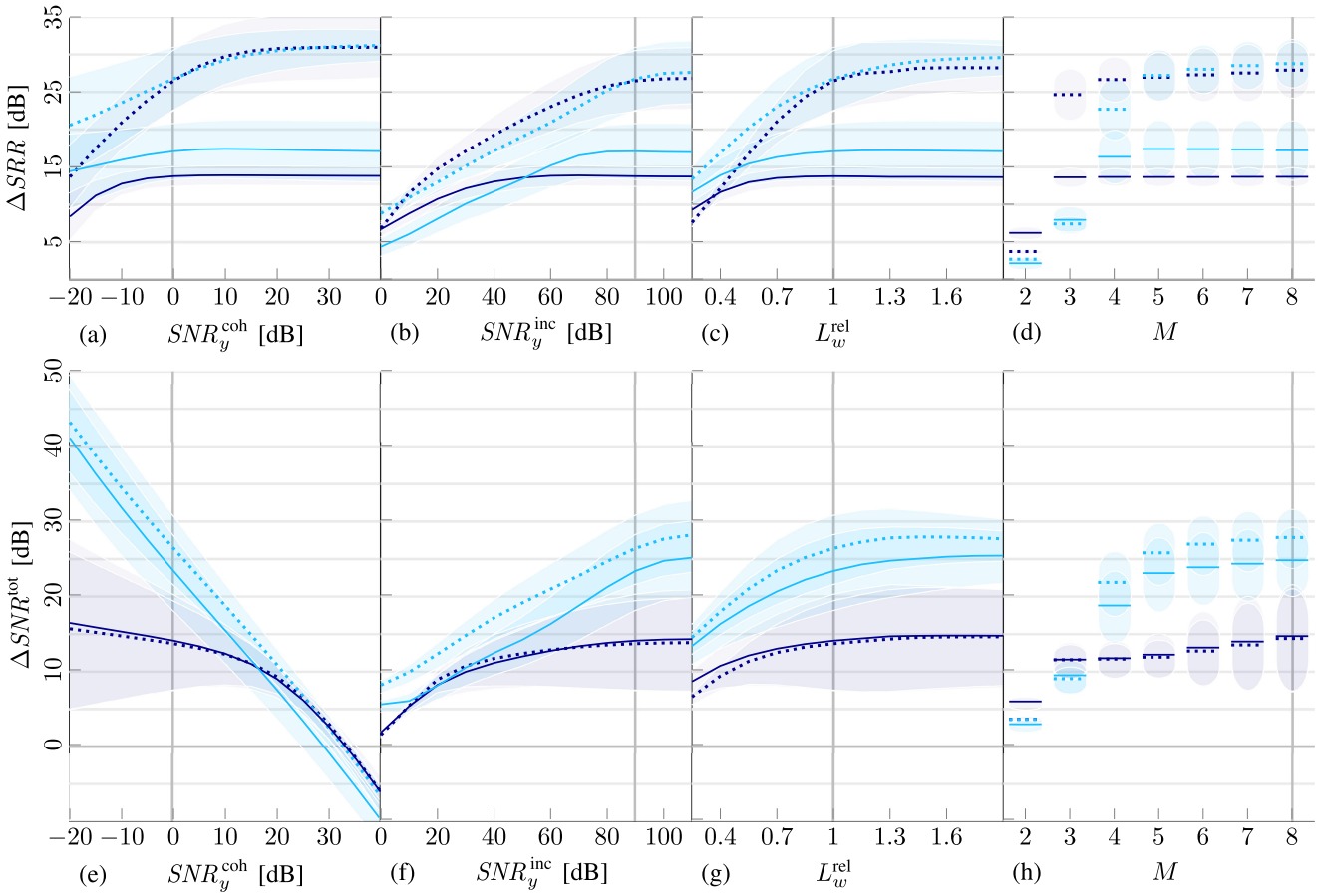


Fig. 6. Dereverberation/noise reduction performance $\Delta SRR/\Delta SNR^{tot}$ versus (a)/(e) the signal-to-coherent-noise ratio SNR_q^{coh}, (b)/(f) the signal-to-incoherent-noise ratio SNR_q^{inc}, (c)/(g) the relative filter length L_w^{rel} and (d)/(h) the number of microphones M for colored and white source signals of the MCLP framework, respectively denoted by [——] and [·····], and the GSC framework, respectively denoted by [——] and [·····]. The vertical grid lines indicate the intersection point of the individual subplots. The shaded areas represent the standard deviation.

both white and colored source signals. MCLP reaches saturation at $M = 3$, while the GSC saturates at $M = 5$ only instead of $M = 4$. This may be caused by remaining low-level incoherent noise and possibly nearly common zeros in the transfer functions corresponding to $\mathbf{H}_B$ in (27) for $M = 4$.

b) Noise reduction: From Fig. 6(e)–(h) we observe that under favorable conditions with predominantly coherent-noise interference, i.e., for low SNR_y^{coh}, high SNR_y^{inc}, $L_w^{\text{rel}} \geq 1$ and sufficiently high M, the GSC [·····, ——] shows increasing improvement in terms of ΔSNR^{tot} for decreasing values of SNR_y^{coh}, while for MCLP [·····, ——], ΔSNR^{tot} is limited to at most 15 dB. the GSC [·····, ——] clearly outperforms MCLP [·····, ——] in terms of ΔSNR^{tot}. This is due to the GSC suppressing the entire coherent-noise component $q_{s_2}(l)$, while MCLP suppresses the late coherent-noise component $\ddot{q}_{s_2}(l)$ only, cf. (47)–(49), (55)–(57) and Table I. Note that MCLP exhibits a stronger standard deviation in ΔSNR^{tot} than the GSC. This is caused by the varying power of the early coherent-noise component $\dot{q}_{s_2}(l)$, as the power of the individual direct components at the output of the MF may be distributed over a range potentially larger than d, depending on the angle between the speech source and the coherent-noise source. In all conditions, the GSC performs somewhat worse for colored signals than white signals, while no significant difference is found for MCLP.

Fig. 6(e) indicates that the GSC exceeds MCLP for $SNR_y^{coh} < 20$ dB, while both frameworks perform similarly for high SNR_y^{coh} values. For the GSC, ΔSNR^{tot} decreases at a rate of slightly less than 10 dB ΔSNR^{tot} per 10 dB SNR_y^{coh}, such that the noise power at the output is almost constant throughout the simulated range. This implies that for $SNR_y^{coh} \geq 35$ dB, the total noise power is in fact even boosted as compared to the output of the MF, both for MCLP and the GSC. Again, this effect can be explained by the limited number of observations L_{obs} in the LS estimate (36), here causing LS to focus on reverberant-speech suppression.

As can be seen from Fig. 6(f), for both white and colored source signals, the GSC exceeds MCLP for higher SNR_y^{inc} values, while the difference reduces for lower values.

As shown in Fig. 6(g), for both white and colored source signals, both MCLP and the GSC again saturate for $L_w^{\text{rel}} \geq 1$. Again, undermodeling does not appear to be extremely critical.

From Fig. 6(h), for both white and colored source signals, we once more find a sharp performance drop for both MCLP and the GSC if $M < 3$ and $M < 4$, respectively. Again, saturation is reached at $M = 3$ and $M = 5$, respectively.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

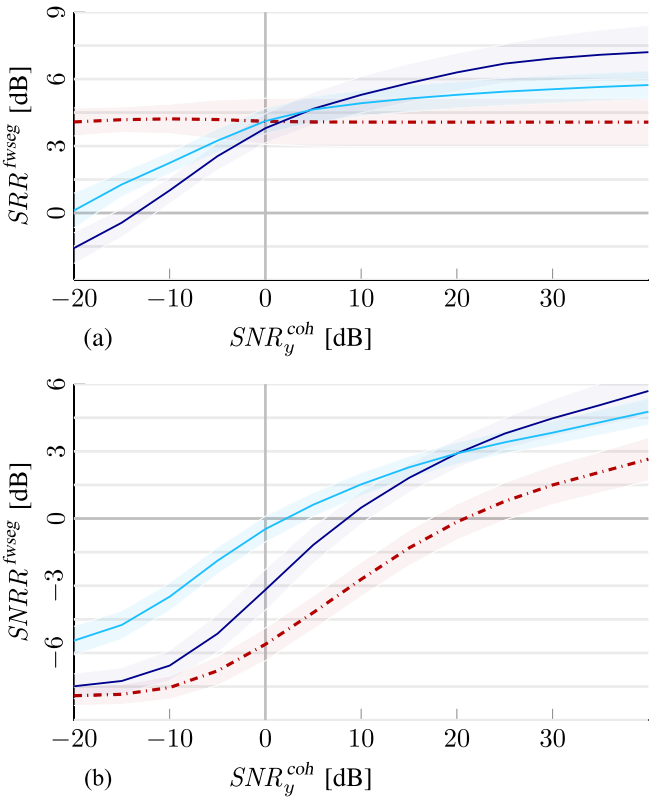


Fig. 7. (a) Dereverberation-only/(b) dereverberation-plus-noise-reduction performance in terms of SRR^{fwseg}/$SNRR^{fwseg}$ versus the signal-to-coherent-noise ratio SNR_y^{coh} of the MF, the MCLP and the GSC framework, respectively denoted by [-·-·-], [——], and [——]. The shaded areas represent the standard deviation.

2) STFT Domain: The performance of both frameworks in terms of SRR^{fwseg} and $SNRR^{fwseg}$ are shown in Fig. 7, where the performance of the MF serves as a reference.

From Fig. 7(a), we note that the dereverberation-only performance of the MF [-·-·-] in terms of SRR^{fwseg} remains almost constant around 4.1 dB. At high SNR_y^{coh} values with predominantly late-reverberant-speech interference, MCLP [——] and the GSC [——] outperform the MF by up to 3.1 dB and 4.1 dB, respectively. Note that in theory, for *complete blocking*, i.e., if (27) is satisfied, the GSC is expected to perform as effectively as MCLP in terms of dereverberation, cf. (47)–(49), (55)–(57), Table I, and the time domain simulation results in Section VII-B1. However, since the sub-band convolution model in (1) poses an approximation of the time-domain convolution only, and since the RETFs used in the GSC blocking matrix are subject to estimation errors, cf. Section VII-A3b, complete blocking is not achieved in the STFT domain. Due to *incomplete blocking*, the GSC hence suffers from some amount of early reverberant-speech cancellation and in addition from incomplete prediction of the late reverberant-speech component $\ddot{q}_{s_1}(l)$, leading to reduced dereverberation performance in comparison to MCLP. At low SNR_y^{coh} values with predominantly coherent-noise interference, we find that both MCLP and the GSC perform worse than the MF, indicating speech distortion. Again, this effect can be explained by the limited number of observations L_{obs} used in the LS estimate (36), causing LS to focus on (late) coherent-noise suppression, cf. Section VII-B1. Since the GSC is able to suppress the early coherent-noise component $\dot{q}_{s_2}(l)$ also, the GSC performance is less affected.

From Fig. 7(b), we note that the dereverberation-plus-noise-reduction performance of the MF in terms of $SNRR^{fwseg}$ ranges between -8 dB for $SNR_y^{coh} = -20$ dB and 2.6 dB for $SNR_y^{coh} = 40$ dB, where the upper limit is still affected by the noise component, as can be seen by comparison to the dereverberation-only performance in Fig. 7(a). At high SNR_y^{coh} values MCLP and the GSC outperform the MF by up to 32.8 dB and 2.1 dB, respectively. At low SNR_y^{coh} values, MCLP performs only somewhat better than the MF, while the GSC in contrast outperforms the MF by up to 5.1 dB. This difference at low SNR_y^{coh} values is expected as MCLP suppresses the late coherent-noise component $\dot{q}_{s_2}(l)$ only, while the GSC suppresses the entire coherent-noise component $q_{s_2}(l)$, cf. (47)–(49), (55)–(57), Table I, and the time domain simulation results in Section VII-B1. Audio examples of the STFT domain simulations are available online [40].

VIII. Conclusion

In this paper, we formally analyzed and compared the MCLP and GSC frameworks in terms of blind dereverberation and noise reduction performance. Both frameworks are theoretically able to perform complete dereverberation if incoherent noise is absent. Due to the use of a blocking matrix, the GSC is theoretically able to completely cancel coherent noise in the absence of incoherent noise, while MCLP cancels the late coherent-noise component only. For complete cancellation, the GSC requires one additional microphone as compared to MCLP. Furthermore, the blocking matrix design requires spatial information in form of the early speech-source RIR or the RETF, which needs to be acquired in practice. In order to confirm the theory and to assess the practical relevance of the theoretical findings, we carried out time domain simulations using oracle knowledge on the early RIRs, resulting in *complete blocking* of the early reverberant-speech component, and STFT domain simulations using estimated REFTs, resulting in *incomplete blocking*.

The simulation results confirm that in terms of noise reduction, as opposed to the GSC performance, the performance of MCLP is limited. In terms of dereverberation, the GSC performs equally well if *complete blocking* is achieved, as expected from the theoretical analysis, but performs inferior for *incomplete blocking*. Both MCLP and the GSC exhibit strong sensitivity to incoherent noise. For both frameworks, dereverberation and noise reduction performance reach their maximum at a relative filter length of about one, while moderate undermodeling of the filter length does not appear to be extremely critical. The simulations further confirm that for one coherent-noise component, the GSC requires four microphones, while MCLP requires three microphones only.

In summary, we can state that if sufficiently many microphones are available and complete blocking is achieved, the GSC performs superior to MCLP in terms of noise reduction and equally well in terms of dereverberation, but inferior in

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

terms of dereverberation for incomplete blocking. In practice therefore, in acoustic conditions with only mild noise but predominantly late-reverberant-speech interference, MCLP is to be preferred, while in case of predominantly noise but mild to moderate late-reverberant-speech interference, the GSC is to be preferred. In acoustic conditions with both strong reverberation and strong noise, combined schemes may be most appropriate.

APPENDIX A

1): Analogously to (16)–(17), let $\dot{\mathbb{C}} \in \mathbb{C}^{(NL_s \times NL_s)}$ and its counterpart $\ddot{\mathbb{C}}$ be defined by

$$\dot{\mathbb{C}} = \begin{pmatrix} \mathbf{I}^{d\times d} & \mathbf{0}^{d\times(NL_s-d)} \\ \mathbf{0}^{(NL_s-d)\times d} & \mathbf{0}^{d\times(NL_s-d)} \end{pmatrix}, \tag{64}$$

$$\ddot{\mathbb{C}} = \mathbf{I}^{NL_s\times NL_s} - \dot{\mathbb{C}}, \tag{65}$$

and let $\underline{\boldsymbol{\Psi}}_{\tilde{s}} \in \mathbb{C}^{NL_s-d\times NL_s-d}$ be the submatrix of $\boldsymbol{\Psi}_{\tilde{s}}$ spanning its last $NL_s - d$ rows and columns, such that

$$\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}} = \begin{pmatrix} \mathbf{0}^{d\times d} & \mathbf{0}^{d\times(NL_s-d)} \\ \mathbf{0}^{(NL_s-d)\times d} & \underline{\boldsymbol{\Psi}}_{\tilde{s}} \end{pmatrix}. \tag{66}$$

With $\mathbf{HB}$ as given in (27), the expression $\left((\mathbf{HB})^H\boldsymbol{\Psi}_{\tilde{s}}\mathbf{HB}\right)^+$ in (54) can then be written as

$$\begin{aligned}\left((\mathbf{HB})^H\boldsymbol{\Psi}_{\tilde{s}}\mathbf{HB}\right)^+ &= (\mathbf{H}_B^H\underline{\boldsymbol{\Psi}}_{\tilde{s}}\mathbf{H}_B)^+ \\ &= \mathbf{H}_B^+\underline{\boldsymbol{\Psi}}_{\tilde{s}}^{-1}\mathbf{H}_B^{+T}.\end{aligned} \tag{67}$$

Inserting (27) and (67) in (54) while noting that $\mathbf{H}_B\mathbf{H}_B^+ = \mathbf{I}$ since $\mathbf{H}_B$ is assumed to have full row rank, and further using (66) and Lemma 1 from Appendix B, we obtain

$$\begin{aligned} z(l) &= \left(\mathbf{HB}\left((\mathbf{HB})^H\boldsymbol{\Psi}_{\tilde{s}}\mathbf{HB}\right)^+(\mathbf{HB})^H\boldsymbol{\Psi}_{\tilde{s}}\mathbf{Hg}\right)^H\mathbf{s}(l) \\ &= \left(\begin{pmatrix} \mathbf{0}^{d\times d} & \mathbf{0}^{d\times(NL_s-d)} \\ \mathbf{0}^{(NL_s-d)\times d} & \underline{\boldsymbol{\Psi}}_{\tilde{s}}^{-1}\end{pmatrix}\boldsymbol{\Psi}_{\tilde{s}}\mathbf{Hg}\right)^H\mathbf{s}(l) \\ &= \left((\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\boldsymbol{\Psi}_{\tilde{s}}\mathbf{Hg}\right)^H\mathbf{s}(l). \end{aligned} \tag{68}$$

With $\dot{\mathbb{C}} + \ddot{\mathbb{C}} = \mathbf{I}$, the matrix $\boldsymbol{\Psi}_{\tilde{s}}$ may be written as

$$\begin{aligned}\boldsymbol{\Psi}_{\tilde{s}} &= \dot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}} + \ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}} \\ &= \dot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}} + \ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}} + \ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\dot{\mathbb{C}}.\end{aligned} \tag{69}$$

Substituting (69) in (68), we find $(\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+(\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}}) = \ddot{\mathbb{C}}$ from (66), while $(\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\dot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}} = \mathbf{0}$, such that $(\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\boldsymbol{\Psi}_{\tilde{s}}$ in (68) becomes

$$(\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\boldsymbol{\Psi}_{\tilde{s}} = \ddot{\mathbb{C}} + (\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\dot{\mathbb{C}}. \tag{70}$$

Using (64)–(65), (16)–(17), and Lemma 1 from Appendix B, the term $(\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\dot{\mathbb{C}}$ in (70) takes the form

$$\begin{aligned}&(\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\dot{\mathbb{C}} \\ &= \begin{pmatrix} (\ddot{\mathbf{C}}\boldsymbol{\Psi}_{\tilde{s}_1}\ddot{\mathbf{C}})^+\ddot{\mathbf{C}}\boldsymbol{\Psi}_{\tilde{s}_1}\dot{\mathbf{C}} & \mathbf{0}^{L_s\times(N-1)L_s} \\ \mathbf{0}^{(N-1)L_s\times L_s} & \mathbf{0}^{(N-1)L_s\times(N-1)L_s}\end{pmatrix}.\end{aligned} \tag{71}$$

Inserting (70) in (68), multiplying out, using (64)–(65) and (14)–(17), it can be shown that

$$\begin{aligned} z(l) &= (\ddot{\mathbb{C}}\mathbf{Hg})^H\mathbf{s}(l) + \left((\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\ddot{\mathbb{C}})^+\ddot{\mathbb{C}}\boldsymbol{\Psi}_{\tilde{s}}\dot{\mathbb{C}}\mathbf{Hg}\right)^H\mathbf{s}(l) \\ &= \sum_{n=2}^{N} q_{s_n}(l) + \ddot{q}_{s_1}(l) \\ &\quad + \left((\ddot{\mathbf{C}}\boldsymbol{\Psi}_{\tilde{s}_1}\ddot{\mathbf{C}})^+\ddot{\mathbf{C}}\boldsymbol{\Psi}_{\tilde{s}_1}\dot{\mathbf{C}}\mathbf{H}_1\mathbf{g}\right)^H\mathbf{s}_0(l). \end{aligned} \tag{72}$$

With Lemma 1 from Appendix B and $\ddot{\mathbf{C}}_d$ defined in (19), $(\ddot{\mathbf{C}}_0\boldsymbol{\Psi}_{\tilde{s}_0}\ddot{\mathbf{C}}_0)^+$ can be written as

$$(\ddot{\mathbf{C}}_1\boldsymbol{\Psi}_{\tilde{s}_1}\ddot{\mathbf{C}}_1)^+ = \ddot{\mathbf{C}}_d^H(\ddot{\mathbf{C}}_d\boldsymbol{\Psi}_{\tilde{s}_1}\ddot{\mathbf{C}}_d^H)\ddot{\mathbf{C}}_d, \tag{73}$$

Substituting (73) into (72) and extracting $\ddot{\mathbf{C}}_d^H$ on the left, we obtain (55)–(56).

2): With Lemma 1 in Appendix B, the first term in (56), $(\ddot{\mathbf{C}}_d\boldsymbol{\Psi}_{\tilde{s}_0}\ddot{\mathbf{C}}_d^H)^+$, can be written as

$$(\ddot{\mathbf{C}}_d\boldsymbol{\Psi}_{\tilde{s}_1}\ddot{\mathbf{C}}_d^H)^+ = \begin{pmatrix} \underline{\boldsymbol{\Psi}}_{\tilde{s}_1}^{-1} & \mathbf{0}^{(L_s-d)\times d} \\ \mathbf{0}^{d\times(L_s-d)} & \mathbf{0}^{d\times d}\end{pmatrix}, \tag{74}$$

where $\underline{\boldsymbol{\Psi}}_{\tilde{s}_1} \in \mathbb{C}^{L_s-d\times L_s-d}$ is the submatrix of $\boldsymbol{\Psi}_{\tilde{s}_1}$ spanning the last $L_s - d$ rows and columns, matching the first term in (48) for $n = 1$. Note that for $d = L_b$, $\underline{\boldsymbol{\Psi}}_{\tilde{s}_1}^{-1} \in \mathbb{C}^{L_s-d\times L_s-d}$ in (74) and $\boldsymbol{\Psi}_{\tilde{s}_1}^{-1} \in \mathbb{C}^{L_s\times L_s}$ in (48) also correspond in terms of dimensions: inserting (26) into (9) yields $L_s - d = L_h + L_w - 1$ in the GSC, while inserting (21) into (9) yields $L_s = L_h + L_w - 1$ in MCLP. Finally, since the second term in (56), $\ddot{\mathbf{C}}_d\boldsymbol{\Psi}_{\tilde{s}_1}\dot{\mathbf{C}}\mathbf{H}_1\mathbf{g}$, is equivalent to the second term in (48), both expressions (56) and (48) yield the same bias component.

APPENDIX B

Lemma 1: The pseudoinverse $\mathbf{A}^+$ of a block-diagonal matrix $\mathbf{A}$ defined by the blocks $\mathbf{A}_n$, $n = 1\ldots N$, on its diagonal is given by a block-diagonal matrix composed of the pseudoinverses $\mathbf{A}_n^+$ of the individual blocks, i.e.,

$$\begin{aligned} &if \qquad \mathbf{A} = \text{blkdiag}\big[\,\mathbf{A}_1,\ \ldots,\ \mathbf{A}_N\,\big], \\ &then \qquad \mathbf{A}^+ = \text{blkdiag}\big[\,\mathbf{A}_1^+,\ \ldots,\ \mathbf{A}_N^+\,\big]. \end{aligned}$$

This lemma can be proven easily by verifying the four criteria defining the pseudoinverse $\mathbf{A}^+$ of the matrix $\mathbf{A}$, i.e., $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$, $(\mathbf{A}\mathbf{A}^+)^H = \mathbf{A}\mathbf{A}^+$, and $(\mathbf{A}^+\mathbf{A})^H = \mathbf{A}^+\mathbf{A}$. It is further important to note that the pseudoinverse of a zero matrix is equal to its transpose.

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] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[3] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, vol. 2007, Dec. 2007, Art. no. 034013.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.

[4] F. Lim and P. Naylor, "Statistical modelling of multichannel blind system identification errors," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Juan-les-Pins, France, Sep. 2014, pp. 119–123.
[5] M. R. P. Thomas, N. D. Gaubitch, E. A. P. Habets, and P. A. Naylor, "An insight into common filtering in noisy SIMO blind system identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 521–524.
[6] M. Triki and D. T. M. Slock, "Blind dereverberation of a single source based on multichannel linear prediction," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Eindhoven, The Netherlands, Sep. 2005, pp. 173–176.
[7] 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.
[8] 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.
[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Aug. 2010.
[10] 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.
[11] 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.
[12] T. Yoshioka, "Dereverberation for reverberation-robust microphone arrays," in *Proc. 21st Eur. Signal Process. Conf.*, Marrakech, Morocco, Sep. 2013, pp. 1–5.
[13] 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, Jun. 2015.
[14] 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.
[15] 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.
[16] 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, pp. 284–288.
[17] 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.
[18] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing persepective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1063, Mar. 2007.
[19] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.
[20] F. Lim, M. R. P. Thomas, and P. Naylor, "MINTFormer: A spatially aware channel equalizer," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.
[21] 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.
[22] 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, QLD, Australia, Apr. 2015, pp. 106–110.
[23] T. Dietzen *et al.*, "Speech dereverberation by data-dependent beamforming with signal pre-whitening," in *Proc. 23rd Eur. Signal Process. Conf.*, Nice, France, Aug. 2015, pp. 2461–2465.
[24] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "On the relation between data-dependent beamforming and multichannel linear prediction for dereverberation," in *Proc. AES 60th Int. Conf. Dereverberation Reverberation Audio, Music, Speech*, Leuven, Belgium, Jan. 2016, pp. 1–8.
[25] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 680–693, Apr. 2016.
[26] M. Delcroix *et al.*, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 2015, pp. 1–15, Dec. 2015.
[27] 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.
[28] 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.
[29] 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.
[30] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. New York, NY, USA: Springer, Jul. 2010.
[31] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. AP-1, no. 30, pp. 27–34, Jan. 1982.
[32] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
[33] 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.
[34] C. D. Meyer, Ed.,*Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA, USA: Soc. Ind. Appl. Math., 2000.
[35] 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 Commun. Microphone Arrays*, San Francisco, CA, USA, Mar. 2017, pp. 116–120.
[36] 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.
[37] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 774–786, Apr. 2015.
[38] Bang and Olufsen, "Music for Archimedes," Compact Disc B&O, 1992.
[39] 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.
[40] T. Dietzen, "Audio examples for IEEE/ACM TASLP 2018," Aug. 2018. [Online]. Available: ftp://ftp.esat.kuleuven.be/pub/SISTA/tdietzen/reports/taslp18/audio



Thomas Dietzen received the Dipl.-Ing. degree from the University of Kaiserslautern, Kaiserslautern, Germany, in 2011. He is currently working toward the Ph.D. degree with KU Leuven, Leuven, Belgium.

From 2012 to 2014, he was a Research Assistant with the University of Heidelberg, Heidelberg, Germany, and with Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany. From 2014 to 2017, he was a Doctoral Researcher with the NXP Semiconductors Belgium NV, Leuven, Belgium, in the frame of the FP7-PEOPLE Marie Curie ITN "Dereverberation and Reverberation of Audio, Music, and Speech." His research is focused on signal processing algorithms for microphone-array-based speech enhancement, specifically on dereverberation and noise reduction. He was a Reviewer for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and the IEEE SIGNAL PROCESSING LETTERS.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.



Ann Spriet received the M.Sc. and Ph.D. degrees in electrical engineering from KU Leuven, Leuven, Belgium, in 1999 and 2004, respectively. From 2004 to 2010, she was a Postdoctoral Researcher with KU Leuven. During her Ph.D. and postdoctoral research, she worked on signal processing algorithms for speech enhancement in hearing aids and cochlear implants. Her main activities are the development of speech enhancement algorithms for mobile devices and technology scouting. She was the recipient of the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control and the International Conference on Acoustics, Speech, and Signal Processing, in 2003 and 2005, respectively. From 2008 to 2010, she was a nominated Officer of EURASIP, where she was active as Newsletter Editor. Since 2010, she has been a Senior Audio Research Engineer with the NXP Semiconductors Belgium NV, Leuven, Belgium. Since 2013, she has been a member of the IEEE Audio and Acoustic Signal Processing Technical Committee.



Wouter Tirry received the M.Sc. degree in physics and the Ph.D. degree in solar physics from the University of Leuven, Leuven, Belgium, in 1994 and 1998, respectively. As a Postdoc, he further pursued his research at the National Centre for Atmospheric Research, Boulder, CO, USA. Since 1999, he has been building up expertise in the domain of speech enhancement for mobile devices with Philips and NXP, Leuven, Belgium, as Research Engineer and System Architect. He is currently a Senior Principal with the Product Line Voice and Audio Solutions, NXP, leading the speech technology development activities.



Simon Doclo (S'95–M'03–SM'13) received the M.Sc. degree in electrical engineering and Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1997 and 2003, respectively. 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, Hamilton, ON, Canada. From 2007 to 2009, he was a Principal Scientist with the Sound and Acoustics Group, NXP Semiconductors, Leuven, Belgium. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and a Scientific Advisor with the Project Group Hearing, Speech, and Audio Technology, 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. He was the recipient of the Master Thesis Award of the Royal Flemish Society of Engineers in 1997, the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003 (with M. Moonen), and the IEEE Signal Processing Society 2008 Best Paper Award (with J. Chen, J. Benesty, A. Huang). He is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Special Area Team 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, CRC Hearing Acoustics). He was the Technical Program Chair of 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, Elsevier *Signal Processing*) and is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and *EURASIP Journal on Advances in Signal Processing*.



Marc Moonen (M'94–SM'06–F'07) is a Full Professor with the Electrical Engineering Department, KU Leuven, Leuven, Belgium, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL, and audio signal processing. He is a Fellow of European Association for Signal Processing (EURASIP) (2018). He was the recipient of the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with P. Vandaele), the 2004 Alcatel Bell (Belgium) Award (with R. 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 G. Leus and with D. Giacobello) and from Elsevier *Signal Processing* (with S. Doclo). He was the Chairman of the IEEE Benelux Signal Processing Chapter (1998–2002), a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and President of EURASIP (2007–2008 and 2011–2012). He was the Editor-in-Chief for the *EURASIP Journal on Applied Signal Processing* (2003–2005), Area Editor for Feature Articles in IEEE SIGNAL PROCESSING MAGAZINE (2012–2014), and has been a member of the editorial board of *Signal Processing*, 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*.



Toon van Waterschoot (S'04–M'12) received the M.Sc. and Ph.D. degrees in electrical engineering, in 2001 and 2009, respectively, both 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 has been serving as 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 *Signal Processing*. He is a Director of the European Association for Signal Processing (EURASIP), a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, and a Founding Member of the EAA Technical Committee in Audio Signal Processing. He was the General Chair of the 60th AES International Conference, Leuven, Belgium (2016), and has been serving on the organizing committee of the European Conference on Computational Optimization (2016) and the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2017). He is a member of EURASIP, ASA, and AES.

Authorized licensed use limited to: KU Leuven Libraries. Downloaded on March 26,2020 at 13:14:35 UTC from IEEE Xplore. Restrictions apply.