Multi-Speaker LCMV Beamformer and Postfilter for Source Separation and Noise Reduction

O. Schwartz, *Student Member, IEEE*, S. Gannot, *Senior Member, IEEE,* and Emanuël A.P. Habets, *Senior Member, IEEE*

***Abstract*—The problem of source separation and noise reduction using multiple microphones is addressed. The minimum mean square error (MMSE) estimator for the multi-speaker case is derived and a novel decomposition of this estimator is presented. The MMSE estimator is decomposed into two stages: i) a *multi-speaker* linearly constrained minimum variance (LCMV) beamformer (BF), and ii) a subsequent *multi-speaker* Wiener postfilter. The first stage separates and enhances the signals of the individual speakers by utilizing the spatial characteristics of the speakers (as manifested by the respective acoustic transfer functions (ATFs)) and the noise spatial correlation matrix, while the second stage exploits the speakers' power spectral density matrix to reduce the residual noise at the output of the first stage. The output vector of the *multi-speaker* LCMV BF is proven to be the *sufficient statistic* for estimating the marginal speech signals in both the classic sense and the Bayesian sense. The log spectral amplitude estimator for the multi-speaker case is also derived given the *multi-speaker* LCMV BF outputs. The performance evaluation was conducted using measured ATFs and directional noise with various signal-to-noise ratio levels. It is empirically verified that the multi-speaker postfilters are beneficial in terms of signal-to-interference plus noise ratio improvement when compared with the single-speaker postfilter.**

I. Introduction

Speech enhancement techniques, utilizing microphone arrays, have attracted the attention of many researchers during the last three decades, especially in the context of hands-free communication tasks. Usually, the received speech signals are contaminated by interfering sources, such as competing speakers and noise sources. Whereas single microphone algorithms might show satisfactory results in noise reduction, they have been found to perform poorly in the presence of one or more competing speakers, as they cannot exploit the spatial diversity exploited by multi-microphone algorithms.

A generalization of the minimum variance distortionless response (MVDR) beamformer (BF), which deals with multiple linear constraints, is the linearly constrained minimum variance (LCMV) BF [1], [2]. The LCMV BF can be applied to construct a beam-pattern, satisfying multiple constraints for a set of directions, while minimizing the output noise power. In [3], [4], the authors proved that the LCMV has an equivalent generalized sidelobe canceller (GSC) structure, which decouples the constraining and minimization operations, as well. In [5], the LCMV was reformulated based on the acoustic transfer functions (ATFs) (or its respective relative transfer functions (RTFs)), rather than based on the delay of the direct-paths. The authors also presented a method to estimate the RTFs, based on the generalized eigenvalue decomposition (GEVD) of the power spectral density (PSD) matrices of the received signals and the stationary noise.

The use of a postfilter is often beneficial to improve the noise reduction performance, especially in nondirectional and nonstationary noise environments [6]. In [7], [8], the authors show that the minimum mean square error (MMSE) estimator of a single speech signal can be equivalently decomposed into two stages, an MVDR-BF and a postfilter. The MVDR-BF exploits the spatial diversity to preserve a distortionless response while minimizing the output noise power. The postfilter is a single-channel Wiener filter [9] that reduces the residual noise at the output of the MVDR-BF by using the PSD of the anechoic speech and the residual noise. Usually, the desired speech PSD is not known in advance. However, in stationary noise environments, the noise PSD can be estimated using speech-absent segments.

Many papers adopted the aforementioned decomposition (i.e., MVDR-BF followed by a single channel postfilter) and proposed various methods to estimate its parameters. In [10], the author presented a practical estimation of the postfilter, based on the entries of the estimated PSD matrix of the observations. This postfilter is only suitable for spatially white noise fields. In [11], the technique was generalized to deal with an arbitrary noise fields, using prior knowledge of the spatial coherence matrix of the noise. In [12], the noise PSD and the speech PSD at the output of the BF were estimated separately, circumventing the overestimation problem encountered in [11]. In [6], [13], [14], a postfilter dealing with nonstationary noise sources was proposed. The authors derived an estimator for the speech presence probability used for estimating the noise PSD at the output of the BF stage. The optimally modified log spectral amplitude (OMLSA) estimator [15] was used as a postfilter using the a priori SNR at the output of the BF. The a priori SNR was recursively estimated using the decision-directed approach [16]. In our previous work [17], the aforementioned decomposition was utilized to jointly suppress reverberation and noise for single-speaker scenarios. The reverberation and the noise were suppressed both in the BF stage and in the postfiltering stage.

In [18], the authors proved that the output of the MVDR-BF is a sufficient statistic for estimating a single speech signal from multichannel inputs in the presence of additive Gaussian

Ofer Schwartz and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: ofer.shwartz@live.biu.ac.il, Sharon.Gannot@biu.ac.il).

E. A. P. Habets is with the International Audio Laboratories Erlangen (a joint institution between the University of Erlangen-Nuremberg and Fraunhofer IIS), Erlangen, Germany (e-mail: emanuel.habets@audiolabs-erlangen.de).

noise. Hence, any MMSE estimator necessarily includes an initial MVDR-BF stage. Furthermore, using this result, the authors were able to generalize the spectral amplitude (SA) [16] and the log spectral amplitude (LSA) [19] estimators to the multichannel case. The SA and LSA estimators are known to be very effective in enhancing noisy speech and significantly improving its quality. The original SA and LSA estimators were derived for the single-channel and single-speaker case. Since the MVDR-BF maps the multichannel input to a single output, the authors in [18] show that the SA and the LSA estimators can be applied at the MVDR output.

The aim of this paper is to extend the above results to the multiple-speakers case that is commonly encountered in cocktail party scenarios [20], [21]. One possibility is to estimate each desired speaker individually by applying the aforementioned single-speaker approach and by considering the other speakers plus ambient noise as interference (as done for example in [22]). In this work, we take a different path, and design a beamformer to simultaneously extract multiple desired speakers. We propose to solve this task by applying the multichannel MMSE estimator of the desired speakers.

Inspired by the decomposition for the single-speaker case presented in e.g. [9], we show that the multichannel Wiener filter (MCWF) can be decomposed into a multi-speaker LCMV BF, followed by a multi-speaker Wiener postfilter. The output of the first stage is a vector which dimensions equal to the number of desired speakers, with each element dominated by a single speech source plus unavoidable residual noise. The multi-speaker Wiener postfilter is a square matrix that minimizes the residual noise at the output signals of the BF stage. In the current work, the output signals of the multi-speaker LCMV BF are proven to be the sufficient statistic (in both the classical and the Bayesian sense) of our estimation problem. The next step is the derivation of an LSA estimator for the multichannel and multi-speaker case. Since the LSA is a Bayesian estimator, the LSA may use only the sufficient statistic rather than all microphone signals. Using the LCMV BF outputs, the marginal sufficient statistic for estimating each speaker signal is defined. The LSA estimator is then derived for each speaker separately, however using all the LCMV BF outputs. The multi-speaker PSD matrix, required for the various proposed estimators, is estimated using a modified decision-directed approach.

The MCWF in the multi-source case was already analyzed in [23] and decomposed into two separable components. The first component, was proved in [24] to be a sufficient statistic (in the classical sense) for estimating concurrent speech signals from the multichannel inputs in the presence of additive Gaussian noise. However, unlike the decomposition in [7], [8] and the new decomposition proposed in this article, the two components in [23] do not constitute a standard spatial filter followed by standard Wiener filter. The proposed decomposition demonstrates several important advantages, most notably the ability to substitute the Wiener posfiltering stage by the LSA estimator, which is advantageous in speech processing.

This paper is organized as follows. In Sec. II, the multichannel and multi-speaker problem is formulated. In Sec. III, the decomposition of the MCWF into the multi-speaker LCMV BF followed by a multi-speaker Wiener postfilter is presented. In Sec. III-D, the output of the multi-speaker LCMV BF is proven to be sufficient statistic for estimating any nonlinear function of the sources. In Sec. IV, the SA and LSA estimators for the multi-speaker and multichannel case are derived. In Sec. V, some practical considerations are given; the multi-speaker LCMV BF is implemented using the GSC approach; and the PSD matrix of the speech signals is estimated using the decision-directed approach. In Sec. VI, the performance of the proposed algorithms is evaluated. Section VII is dedicated to concluding remarks.

II. Problem Formulation

The source separation and enhancement problem is formulated in the short-time Fourier transform (STFT) domain with ℓ denoting the frame index and k denoting the frequency index. Assume that the sound of J coherent speakers are captured by N microphones. The i-th microphone signal can be expressed as

$$Y_i(\ell,k) = \sum_{j=1}^{J} X_{i,j}(\ell,k) + V_i(\ell,k),\ i = 1,2,\ldots,N \quad (1)$$

where $X_{i,j}(\ell,k)$ denotes the speech signal of the jth speaker as received by the ith microphone and $V_i(\ell,k)$ denotes the ambient and sensor noise.

We also assume that the observed speech, as received by the ith microphone, can be approximated in the STFT domain as a multiplication of an anechoic speech signal $S_j(\ell,k)$ with a time-invariant ATF $G_{i,j}(k)$ (i.e., assuming a static scenario) relating the speaker position and the ith microphone, i.e.

$$X_{i,j}(\ell,k) = G_{i,j}(k) S_j(\ell,k). \quad (2)$$

The N microphone signals can be stacked in a vector form

$$\begin{aligned}\mathbf{y}(\ell,k) &= \left[\begin{array}{cccc} Y_1(\ell,k) & Y_2(\ell,k) & \ldots & Y_N(\ell,k) \end{array}\right]^{\mathrm{T}} \\ &= \sum_{j=1}^{J} \mathbf{x}_j(\ell,k) + \mathbf{v}(\ell,k) \\ &= \sum_{j=1}^{J} \mathbf{g}_j(k) S_j(\ell,k) + \mathbf{v}(\ell,k) \\ &= \mathbf{G}(k)\mathbf{s}(\ell,k) + \mathbf{v}(\ell,k), \end{aligned} \quad (3)$$

where

$$\begin{aligned}\mathbf{x}_j(\ell,k) &= \left[\begin{array}{cccc} X_{1,j}(\ell,k) & X_{2,j}(\ell,k) & \cdots & X_{N,j}(\ell,k) \end{array}\right]^{\mathrm{T}} \\ \mathbf{v}(\ell,k) &= \left[\begin{array}{cccc} V_1(\ell,k) & V_2(\ell,k) & \cdots & V_N(\ell,k) \end{array}\right]^{\mathrm{T}} \\ \mathbf{g}_j(k) &= \left[\begin{array}{cccc} G_{1,j}(k) & G_{2,j}(k) & \cdots & G_{N,j}(k) \end{array}\right]^{\mathrm{T}} \\ \mathbf{G}(k) &= \left[\begin{array}{cccc} \mathbf{g}_1(k) & \mathbf{g}_2(k) & \cdots & \mathbf{g}_J(k) \end{array}\right] \\ \mathbf{s}(\ell,k) &= \left[\begin{array}{cccc} S_1(\ell,k) & S_2(\ell,k) & \cdots & S_J(\ell,k) \end{array}\right]^{\mathrm{T}}. \end{aligned}$$

The probability density function (p.d.f.) of the observed data given the anechoic speech signal is modelled as a complex-Gaussian

$$\begin{aligned} f(\mathbf{y}(\ell,k)|\mathbf{s}(\ell,k);\mathbf{G}(k),\mathbf{\Phi}_{\mathbf{v}}(k)) = \\ \mathcal{N}_{\mathrm{C}}\left(\mathbf{y}(\ell,k);\mathbf{G}(k)\mathbf{s}(\ell,k),\mathbf{\Phi}_{\mathbf{v}}(k)\right), \end{aligned} \quad (4)$$

where $\mathbf{\Phi}_{\mathbf{v}}(k) = E\{\mathbf{v}(\ell,k)\mathbf{v}^{\mathrm{H}}(\ell,k)\}$ is the PSD matrix of the ambient noise and $\mathcal{N}_{\mathrm{C}}(\cdot;\cdot,\cdot)$ denotes the complex Gaussian probability

$$\mathcal{N}_{\mathrm{C}}(\mathbf{x};\boldsymbol{\mu},\mathbf{\Phi}) = \frac{1}{\pi^N \det(\mathbf{\Phi})} \exp\left((\mathbf{x}-\boldsymbol{\mu})^{\mathrm{H}} \mathbf{\Phi}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right), \tag{5}$$

where $\mathbf{x}$ is a Gaussian random vector, $\boldsymbol{\mu}$ is its mean and $\mathbf{\Phi}$ is its PSD matrix. The p.d.f. of the anechoic speech signals is modelled by:

$$f(\mathbf{s}(\ell,k);\mathbf{\Phi}_{\mathbf{s}}(\ell,k)) = \mathcal{N}_{\mathrm{C}}(\mathbf{s}(\ell,k);\mathbf{0},\mathbf{\Phi}_{\mathbf{s}}(\ell,k)) \tag{6}$$

where $\mathbf{\Phi}_{\mathbf{s}}(\ell,k) = E\{\mathbf{s}(\ell,k)\mathbf{s}^{\mathrm{H}}(\ell,k)\}$ is the PSD matrix of the multiple speakers.

The aim of this work is to provide an optimal (in the MMSE sense) multichannel estimate of a *filtered version* of the J speaker signals

$$\mathbf{s}_{\mathrm{F}}(\ell,k) = \begin{bmatrix} F_1(k)S_1(\ell,k) & F_2(k)S_2(\ell,k) & \cdots & F_J(k)S_J(\ell,k) \end{bmatrix}^{\mathrm{T}}. \tag{7}$$

For example, when using $F_j(k) = G_{1,j}(k)$ (as proposed in [25]), we aim at estimating the speech of the jth speaker as received by the first microphone,

$$\mathbf{s}_{\mathrm{F}}(\ell,k) = \begin{bmatrix} X_{1,1}(\ell,k) & X_{1,2}(\ell,k) & \cdots & X_{1,J}(\ell,k) \end{bmatrix}^{\mathrm{T}}. \tag{8}$$

In the rest of the paper, the general case in (7) is assumed.

The well-known MMSE estimate of $\mathbf{s}_{\mathrm{F}}(\ell,k)$ given the microphone signals is given by [24, Eq. 7.34]

$$\underset{\hat{\mathbf{s}}_{\mathrm{F}}}{\operatorname{argmin}}\, E\left\{\|\hat{\mathbf{s}}_{\mathrm{F}}(\mathbf{y}(\ell,k)) - \mathbf{s}_{\mathrm{F}}(\ell,k)\|^2\right\} = E\left\{\mathbf{s}_{\mathrm{F}}(\ell,k)|\mathbf{y}(\ell,k)\right\}. \tag{9}$$

The next section is dedicated to the derivation of the MMSE estimator of $\mathbf{s}_{\mathrm{F}}(\ell,k)$.

III. Optimal Multichannel Noise Reduction and Speaker Separation

In this section we first describe the optimal MMSE estimator of the filtered signal $\mathbf{s}_{\mathrm{F}}(\ell,k)$. In the following, whenever applicable, the frequency index k and the time index ℓ are omitted for brevity. To simplify the derivation, we first rewrite the received signal model (3) in terms of the filtered signal

$$\begin{aligned}\mathbf{y} &= \sum_{j=1}^{J} \frac{\mathbf{g}_j}{F_j} F_j S_j + \mathbf{v} \\ &= \sum_{j=1}^{J} \tilde{\mathbf{g}}_j S_{\mathrm{F},j} + \mathbf{v} \\ &= \widetilde{\mathbf{G}}\mathbf{s}_{\mathrm{F}} + \mathbf{v},\end{aligned} \tag{10}$$

where $\tilde{\mathbf{g}}_j \equiv \mathbf{g}_j/F_j$ are the normalized ATFs, $S_{\mathrm{F},j} \equiv F_j S_j$ are the filtered speech signals and $\widetilde{\mathbf{G}} = \begin{bmatrix} \tilde{\mathbf{g}}_1 & \tilde{\mathbf{g}}_2 & \cdots & \tilde{\mathbf{g}}_J \end{bmatrix}$.

In Sec. III-A, the proposed decomposition of the MMSE estimator is presented. Later, in Sec. III-B and Sec. III-C, the two components are discussed.

A. MMSE Estimator and its Decomposition

Since $\mathbf{s}_{\mathrm{F}}$ and $\mathbf{y}$ are assumed to be zero-mean complex-Gaussian random variables, the MMSE estimator of $\mathbf{s}_{\mathrm{F}}$ is given by the MCWF [24, Eq. (7.167)]:

$$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{MCWF}} &= E\{\mathbf{s}_{\mathrm{F}}\mathbf{y}^{\mathrm{H}}\} \times E\{\mathbf{y}\mathbf{y}^{\mathrm{H}}\}^{-1}\mathbf{y} \\ &= \mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}\widetilde{\mathbf{G}}^{\mathrm{H}} \times \left[\widetilde{\mathbf{G}}\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}\widetilde{\mathbf{G}}^{\mathrm{H}} + \mathbf{\Phi}_{\mathbf{v}}\right]^{-1}\mathbf{y},\end{aligned} \tag{11}$$

where $\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}$ is the PSD matrix of $\mathbf{s}_{\mathrm{F}}$.

Using the Woodbury identity [26] and some algebraic steps (as shown in [23, Eq. (6.181)]), $\hat{\mathbf{s}}_{\mathrm{MCWF}}$ can be expressed as

$$\hat{\mathbf{s}}_{\mathrm{MCWF}} = \left(\mathbf{I} + \mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}\widetilde{\mathbf{G}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}\right)^{-1}\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}} \times \widetilde{\mathbf{G}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}^{-1}\mathbf{y}, \tag{12}$$

Assuming that $\mathbf{A}$ and $\mathbf{B}$ are invertible matrices, the following identity (adapted from [26, Eq. (167)]) may be used:

$$(\mathbf{I}+\mathbf{A}\mathbf{B})^{-1}\mathbf{A} = \mathbf{A}\left(\mathbf{A}+\mathbf{B}^{-1}\right)^{-1}\mathbf{B}^{-1}. \tag{13}$$

Identifying $\mathbf{A} = \mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}$ and $\mathbf{B} = \widetilde{\mathbf{G}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}$, and applying the above identity to the right-hand side of (12) we obtain:

$$\hat{\mathbf{s}}_{\mathrm{LCMV+MCWPF}} = \underbrace{\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}\left(\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}} + \left(\widetilde{\mathbf{G}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}\right)^{-1}\right)^{-1}}_{\mathbf{H}_{\mathrm{WPF}}^{\mathrm{H}}} \times \underbrace{\left(\widetilde{\mathbf{G}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}\right)^{-1}\widetilde{\mathbf{G}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}^{-1}}_{\mathbf{H}_{\mathrm{LCMV}}^{\mathrm{H}}}\mathbf{y}, \tag{14}$$

where $\mathbf{H}_{\mathrm{LCMV}}$ is an $N \times J$ matrix which denotes the multi-speaker LCMV BF, and $\mathbf{H}_{\mathrm{WPF}}$ is an $J \times J$ symmetric matrix which denotes the multi-speaker Wiener postfilter that is applied to $\mathbf{H}_{\mathrm{LCMV}}^{\mathrm{H}}\mathbf{y}$.

For the decomposition to be valid, $\mathbf{B} = \widetilde{\mathbf{G}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}$ needs to be invertible. This requirement is satisfied only when i) the noise PSD matrix is of full rank, and ii) the column rank of $\mathbf{G}$ equals J, which is true when $N \geq J$ and the RTFs are linearly independent.

The proposed decomposition (14) provides some benefits over the direct implementation (12): i) the filters of the proposed decomposition are well-known and their behaviour in the presence of estimations errors, as well as methods to increase their robustness are well understood, ii) in a static scenario, the LCMV BF is time-invariant, iii) the LCMV output signal vector can be used to estimate the speaker PSD matrix as shown in Sec. V-B, iv) the LCMV BF can be efficiently implemented using GSC structure as shown in Sec V-A, and v) the multi-speaker LSA estimator can be derived using the LCMV BF outputs as derived in Sec. IV. The attenuation of these single-channel filters can be further restricted to mitigate musical noise as discussed in Sec. IV.

Alternative decompositions of (12) exists. According to [23, Eq. (6.182)], the MCWF can be decomposed into J MVDR BFs, each steered to the jth speaker while minimizing the power of the other $J-1$ speakers and the noise, and a subsequent optimal single-channel Wiener filter. It should be noted that even in a static scenarios, the MVDR BFs are time-varying due to the non-stationarity of the $J-1$ speakers, and that all MVDR BFs are different from each

other. In the proposed decomposition, only one (multiple output) LCMV BF is applied, which significantly reduces the computational complexity. Moreover, as the resulting LCMV BF is time-invariant, the computational complexity is even further reduced, and the resulting scheme becomes more robust to unknown environment.

B. The Multi-speaker LCMV BF Criterion

The multi-speaker LCMV $\mathbf{H}_{\text{LCMV}}$, defined in (14), can be also obtained by solving the following optimization criterion

$$\mathbf{H}_{\text{LCMV}} = \underset{\mathbf{H}}{\operatorname{argmin}} \operatorname{Tr}\left[\mathbf{H}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{H}\right] \text{ subject to } \mathbf{H}^{\text{H}}\widetilde{\mathbf{G}} = \mathbf{I}, \quad (15)$$

where $\operatorname{Tr}[\cdot]$ denotes the trace operation. According to this criterion, the multiple speech signals are undistorted (under the definition with F_j) while the noise is minimized. The output of the multi-speaker LCMV is given by

$$\widehat{\mathbf{s}}_{\text{LCMV}} \equiv \mathbf{H}_{\text{LCMV}}^{\text{H}}\mathbf{y} = \mathbf{s}_{\text{F}} + \mathbf{v}_{\text{RE}}, \quad (16)$$

where $\mathbf{v}_{\text{RE}} \equiv \mathbf{H}_{\text{LCMV}}^{\text{H}}\mathbf{v}$ denotes the residual noise signal vector of length J. The solution of the criterion in (15) can be obtained by applying the method of Lagrange multipliers. Denote $\boldsymbol{\Lambda}$ as a $J \times J$ constraint matrix. The total expression to be minimized is

$$\operatorname{Tr}\left[\mathbf{H}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{H}\right] - \operatorname{Tr}\left[\boldsymbol{\Lambda}^{\text{H}}\left(\mathbf{H}^{\text{H}}\widetilde{\mathbf{G}} - \mathbf{I}\right)\right]. \quad (17)$$

Setting to zero the first-order derivative with respect to $\mathbf{H}$ yields

$$\frac{\partial}{\partial \mathbf{H}}\left(\operatorname{Tr}\left[\mathbf{H}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{H}\right] - \operatorname{Tr}\left[\boldsymbol{\Lambda}^{\text{H}}\left(\mathbf{H}^{\text{H}}\widetilde{\mathbf{G}} - \mathbf{I}\right)\right]\right) = 2\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{H} - \widetilde{\mathbf{G}}\boldsymbol{\Lambda}^{\text{H}} = \mathbf{0}. \quad (18)$$

Using the constraint $\mathbf{H}^{\text{H}}\widetilde{\mathbf{G}} = \mathbf{I}$, we obtain the solution for $\boldsymbol{\Lambda}$ and consequently

$$\mathbf{H}_{\text{LCMV}} = \boldsymbol{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}\left(\widetilde{\mathbf{G}}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}\right)^{-1}\mathbf{I}. \quad (19)$$

Note that each column of the multi-speaker LCMV is a standard LCMV [1], [2] steered towards one of the speakers while cancelling all other $J-1$ speakers. Hence, each of the outputs of $\mathbf{H}_{\text{LCMV}}$ should be dominated by solely one speaker. In the next section we prove that $\widehat{\mathbf{s}}_{\text{LCMV}}$ is the *sufficient statistic* (both in the classic sense and in the Bayesian sense) for estimating $\mathbf{s}_{\text{F}}$ from the microphone signals $\mathbf{y}$.

C. Multi-Speaker Wiener Postfilter

The multi-speaker Wiener postfilter $\mathbf{H}_{\text{WPF}}$ can be cast as the multi-speaker MMSE estimator of $\mathbf{s}_{\text{F}}$ given the multi-speaker LCMV outputs $\widehat{\mathbf{s}}_{\text{LCMV}}$ defined in (16)

$$\widehat{\mathbf{s}}_{\text{LCMV+SCWPF}} = E\{\mathbf{s}_{\text{F}}|\widehat{\mathbf{s}}_{\text{LCMV}}\} = \mathbf{H}_{\text{WPF}}^{\text{H}}\widehat{\mathbf{s}}_{\text{LCMV}}. \quad (20)$$

Let us first define the residual noise power at the output of the multi-speaker LCMV BF as

$$\boldsymbol{\Phi}_{\mathbf{v},\text{RE}} \equiv E\left\{\mathbf{v}_{\text{RE}}\mathbf{v}_{\text{RE}}^{\text{H}}\right\} = \mathbf{H}_{\text{LCMV}}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{H}_{\text{LCMV}} = \left(\widetilde{\mathbf{G}}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}\right)^{-1}. \quad (21)$$

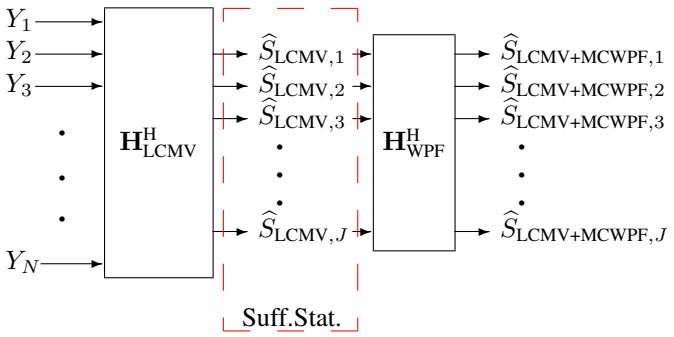


Fig. 1: Block diagram of the proposed decomposition.

Using this definition, the multi-speaker Wiener postfilter in (14) can be rewritten as:

$$\mathbf{H}_{\text{WPF}} = \boldsymbol{\Phi}_{\mathbf{s}_{\text{F}}}\left(\boldsymbol{\Phi}_{\mathbf{s}_{\text{F}}} + \boldsymbol{\Phi}_{\mathbf{v},\text{RE}}\right)^{-1}. \quad (22)$$

Note that $\mathbf{H}_{\text{WPF}}$ is a multi-speaker postfilter that further enhances the speakers in the MMSE sense, but may sacrifice the separation capabilities and may distort the speech signals (equivalent to the single-microphone postfilter that reduces the mean square error (MSE) but distorts the desired speaker at the output of an MVDR beamformer [18]).

D. Sufficient Statistic

In this section, the term $\text{T}(\mathbf{y}) \equiv \widehat{\mathbf{s}}_{\text{LCMV}}$ in (16) is shown to be sufficient statistic for estimating $\mathbf{s}_{\text{F}}$ from the measurements $\mathbf{y}$, in both the *classic sense* and the *Bayesian sense*.

Following [24, Eq. (3.55)], the signal vector $\widetilde{\mathbf{G}}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{y}$ comprises all information required to obtained a maximum likelihood estimate of $\mathbf{s}_{\text{F}}$, i.e. the signal vector $\widetilde{\mathbf{G}}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{y}$ is a sufficient statistic in the classical sense. Since $\widetilde{\mathbf{G}}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\mathbf{y}$ is multiplied by the invertible matrix $\left(\widetilde{\mathbf{G}}^{\text{H}}\boldsymbol{\Phi}_{\mathbf{v}}^{-1}\widetilde{\mathbf{G}}\right)^{-1}$, the LCMV output $\widehat{\mathbf{s}}_{\text{LCMV}}$ is also a sufficient statistic in the classical sense.

According to [27, Theorem 2.14], the sufficiency in the classical sense also implies *Bayesian sufficiency*

$$f(\mathbf{s}_{\text{F}}|\mathbf{y}) = f(\mathbf{s}_{\text{F}}|\text{T}(\mathbf{y})), \quad (23)$$

i.e., the p.d.f. of $\mathbf{s}_{\text{F}}$ given $\mathbf{y}$ equals to the p.d.f. of $\mathbf{s}_{\text{F}}$ given $\text{T}(\mathbf{y})$. The last result implies that the MMSE estimator of any (nonlinear) function of $\mathbf{s}_{\text{F}}$ given the measurements $\mathbf{y}$ can utilize the sufficient statistic $\text{T}(\mathbf{y})$ rather than the original measurements. Let $\boldsymbol{\rho}(\mathbf{s}_{\text{F}})$ be a function of $\mathbf{s}_{\text{F}}$, then the latter argument can be justified using (23) by the following derivation:

$$\begin{aligned} E\{\boldsymbol{\rho}(\mathbf{s}_{\text{F}})|\mathbf{y}\} &= \int\cdots\int \boldsymbol{\rho}(\mathbf{s}_{\text{F}})\, f(\mathbf{s}_{\text{F}}|\mathbf{y})\, \text{d}S_{\text{F},1}\cdots\text{d}S_{\text{F},J} \\ &= \int\cdots\int \boldsymbol{\rho}(\mathbf{s}_{\text{F}})\, f(\mathbf{s}_{\text{F}}|\text{T}(\mathbf{y}))\, \text{d}S_{\text{F},1}\cdots\text{d}S_{\text{F},J} \\ &= E\{\boldsymbol{\rho}(\mathbf{s}_{\text{F}})|\text{T}(\mathbf{y})\}. \end{aligned} \quad (24)$$

A block diagram of the proposed decomposition is depicted in Figure 1, where $\widehat{S}_{\text{LCMV},j}$ and $\widehat{S}_{\text{LCMV+MCWPF},j}$ denote the jth element of $\widehat{\mathbf{s}}_{\text{LCMV}}$ and $\widehat{\mathbf{s}}_{\text{LCMV+MCWPF}}$, respectively.

IV. Multi-speaker LSA Estimator

In this section, the multi-speaker LSA estimator [19] is derived. Note, that the SA estimator [16] can be similarly derived. In this work we focus on the LSA estimator, since it is a more common estimator in speech enhancement applications.

In Sec. III-D, $\widehat{\mathbf{s}}_{\text{LCMV}}$ was proven to be the sufficient statistic for estimating any function of $\mathbf{s}_{\text{F}}$. Accordingly, denoting the amplitude of the jth filtered speaker as $A_j \equiv |S_{\text{F},j}|$, the multi-speaker LSA estimator of $S_{\text{F},j}$ can be defined by

$$\widehat{A}_{\text{LSA},j} = \exp\left(E\left\{\log\left(A_j\right)|\widehat{\mathbf{s}}_{\text{LCMV}}\right\}\right). \tag{25}$$

Since $\widehat{\mathbf{s}}_{\text{LCMV}}$ is a J-dimensional column vector, the LSA estimator cannot be straightforwardly derived by following the original derivation in [19]. In the following, instead of jointly estimating the individual speakers, we aim at estimating only the jth speaker from the sufficient statistic in (16). Note that according to [23, Eq. (6.182)], the estimate of the jth speech signal (that is obtained using the MMSE estimator that jointly estimates the individual speech signals) is identical to the MMSE estimator that provides an estimate of the jth speech signal while treating the other speech signals plus the additive noise as interference. For that, we further reduce $\widehat{\mathbf{s}}_{\text{LCMV}}$ to construct the marginal sufficient statistic for estimating the jth speaker. To derive an MMSE estimator of a nonlinear function of the jth speaker, we may decompose the output vector of the LCMV beamformer by recasting (16) as a summation of a desired speaker and a combination of other speakers and additive noise:

$$\widehat{\mathbf{s}}_{\text{LCMV}} = \mathbf{i}_j\, S_{\text{F},j} + \mathbf{z}_{\bar{j}} \tag{26}$$

where $\mathbf{i}_j$ is the jth column of the identity matrix $\mathbf{I}$ and

$$\mathbf{z}_{\bar{j}} \equiv \left(\mathbf{I} - \text{Diag}\left[\mathbf{i}_j\right]\right)\mathbf{s}_{\text{F}} + \mathbf{v}_{\text{RE}}. \tag{27}$$

Assuming that the speakers are statistically independent, the terms on the right-hand side of (26) are also statistically independent. Thus, according to [18, Eq. (12)], the sufficient statistic for estimating $S_{j,\text{F}}$ from $\widehat{\mathbf{s}}_{\text{LCMV}}$ is

$$\text{T}_j\left(\mathbf{y}\right) = \text{T}_j\left(\widehat{\mathbf{s}}_{\text{LCMV}}\right) \equiv \mathbf{w}_j^{\text{H}}\widehat{\mathbf{s}}_{\text{LCMV}}, \tag{28}$$

where $\mathbf{w}_j$ is the MVDR-BF steered towards the jth speaker while minimizing the power of $\mathbf{z}_{\bar{j}}$

$$\mathbf{w}_j = \frac{\boldsymbol{\Sigma}_{\bar{j}}^{-1}\mathbf{i}_j}{\mathbf{i}_j^{\text{T}}\boldsymbol{\Sigma}_{\bar{j}}^{-1}\mathbf{i}_j}, \tag{29}$$

with

$$\begin{aligned}\boldsymbol{\Sigma}_{\bar{j}} &\equiv E\left\{\mathbf{z}_{\bar{j}}\mathbf{z}_{\bar{j}}^{\text{H}}\right\} \\ &= \left(\mathbf{I} - \text{Diag}\left[\mathbf{i}_j\right]\right)\boldsymbol{\Phi}_{\mathbf{s}_{\text{F}}}\left(\mathbf{I} - \text{Diag}\left[\mathbf{i}_j\right]\right)^{\text{T}} + \boldsymbol{\Phi}_{\mathbf{v},\text{RE}}.\end{aligned} \tag{30}$$

Similarly to (26), the sufficient statistic $\text{T}_j\left(\widehat{\mathbf{s}}_{\text{LCMV}}\right)$ can be represented as

$$\text{T}_j\left(\widehat{\mathbf{s}}_{\text{LCMV}}\right) = S_{\text{F},j} + \bar{Z}_{\text{RE},j}, \tag{31}$$

where $\bar{Z}_{\text{RE},j} \equiv \mathbf{w}_j^{\text{H}}\mathbf{z}_{\bar{j}}$ is the residual interference with variance equal to $E\left\{\left|\bar{Z}_{\text{RE},j}\right|^2\right\} = \left(\mathbf{i}_j^{\text{T}}\boldsymbol{\Sigma}_{\bar{j}}^{-1}\mathbf{i}_j\right)^{-1}$. Utilizing the single-channel model in (31), the multi-speaker LSA estimator can now be stated. Denoting $R_j \equiv |\text{T}_j\left(\widehat{\mathbf{s}}_{\text{LCMV}}\right)|$, the multi-speaker LSA estimator can be stated as [19]:

$$\widehat{A}_{\text{LSA},j} = \exp\left(E\left\{\log\left(A_j\right)|\text{T}_j\left(\widehat{\mathbf{s}}_{\text{LCMV}}\right)\right\}\right) = H_{\text{LSA},j}R_j, \tag{32}$$

where

$$H_{\text{LSA},j} = \max\left(\frac{\xi_j}{1+\xi_j}\exp\left(\frac{1}{2}\int_{\nu_j}^{\infty}\frac{e^{-t}}{t}dt\right), H_{\text{min}}\right) \tag{33}$$

and ν_j, ξ_j and γ_j are defined by

$$\nu_j = \frac{\xi_j}{1+\xi_j}\gamma_j, \tag{34a}$$

$$\xi_j = \frac{\phi_{S_{\text{F},j}}}{\left(\mathbf{i}_j^{\text{T}}\boldsymbol{\Sigma}_{\bar{j}}^{-1}\mathbf{i}_j\right)^{-1}}, \tag{34b}$$

$$\text{and } \gamma_j = \frac{R_j^2}{\left(\mathbf{i}_j^{\text{T}}\boldsymbol{\Sigma}_{\bar{j}}^{-1}\mathbf{i}_j\right)^{-1}}, \tag{34c}$$

where the PSD $\phi_{S_{\text{F},j}}$ is the jth element of $\text{Diag}[\boldsymbol{\Phi}_{\mathbf{s}_{\text{F}}}]$. Practically, to minimize speech distortion and to mitigate *musical noise* [28], the LSA filter in (33) is lower-bounded by a time-and-frequency-dependent gain H_{min}. Since both $\mathbf{H}_{\text{LCMV}}$ and $\mathbf{w}_j$ are distortionless beamformers, the lower-bound is only applied at the LSA stage.

Finally, the LSA estimate of the jth speaker's signal using the multichannel readings is given by

$$\widehat{S}_{\text{LCMV+MCLSA},j} = H_{\text{LSA},j}\text{T}_j\left(\widehat{\mathbf{s}}_{\text{LCMV}}\right). \tag{35}$$

Using (28) and (35), the LSA estimators of all speakers can be constructed as

$$\widehat{\mathbf{s}}_{\text{LCMV+MCLSA}} = \left[\begin{array}{ccc}\widehat{S}_{\text{LCMV+MCLSA},1} & \dots & \widehat{S}_{\text{LCMV+MCLSA},J}\end{array}\right]. \tag{36}$$

The estimator of all desired speakers can be stated in matrix form

$$\widehat{\mathbf{s}}_{\text{LCMV+MCLSA}} = \mathbf{H}_{\text{LSA}}^{\text{H}}\widehat{\mathbf{s}}_{\text{LCMV}}, \tag{37}$$

where

$$\mathbf{H}_{\text{LSA}} = \mathbf{Q}\,\text{Diag}\left[\begin{array}{ccc}H_{\text{LSA},1} & \dots & H_{\text{LSA},J}\end{array}\right] \tag{38}$$

and

$$\mathbf{Q} = \left[\begin{array}{ccc}\mathbf{w}_1 & \dots & \mathbf{w}_J\end{array}\right]. \tag{39}$$

Note that in this implementation, three stages are eventually executed: i) multi-speaker LCMV BF, ii) J concatenated BFs $\mathbf{Q}$, and iii) J LSA filters $H_{\text{LSA},j}$.

To better compare the performance of the multichannel Wiener postfilter in (22) and the multi-speaker LSA postfilter in (37), it is proposed to use the lower-bound H_{min} also with $\mathbf{H}_{\text{WPF}}$ as well, the multi-speaker Wiener postfilter in (22) can be alternatively implemented similarly to the LSA in (37)-(39) as

$$\mathbf{H}_{\text{WPF}} = \mathbf{Q}\,\text{Diag}\left[\begin{array}{ccc}H_{\text{WPF},1} & \dots & H_{\text{WPF},J}\end{array}\right], \tag{40}$$

where

$$H_{\text{WPF},j} = \max\left(\frac{\xi_j}{1+\xi_j}, H_{\text{min}}\right). \tag{41}$$

V. Practical Considerations

We have shown that for any estimator of $\mathbf{s}_{\mathrm{F}}$ (e.g. the MCWF and LSA postfilter estimators), the sufficient statistic $\mathbf{s}_{\mathrm{LCMV}}$ can be precalculated using the multi-speaker LCMV BF. To reduce the computational burden, a GSC version of the multi-speaker LCMV BF can be used. Such GSC version is derived in Sec. V-A. Since the PSD matrix of the speakers $\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}$ is usually unknown in advance, a decision-directed based estimation of $\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}$ is proposed in Sec. V-B. In Sec. V-C, $\mathrm{T}_j\left(\widehat{\mathbf{s}}_{\mathrm{LCMV}}\right)$ is calculated using the GSC structure as well to avoid inversion of matrices with low condition-number.

A. GSC implementation of the Multi-Speaker LCMV BF

A GSC structure [25], [29] may be used to efficiently implement the LCMV beamformer. In our case, there is a significant advantage in implementing $\mathbf{H}_{\mathrm{LCMV}}$ in a GSC structure, since an inversion of an $(N-J)\times(N-J)$ matrix is required rather than an inversion of an $N\times N$ matrix in the original closed-form LCMV implementation (19). By following the GSC formulation, the multi-speaker LCMV beamformer can be written as

$$\mathbf{H}_{\mathrm{LCMV}}=\mathbf{H}_0-\mathbf{B}\mathbf{H}_{\mathrm{NC}}, \tag{42}$$

where $\mathbf{H}_0$ is the fixed multi-speaker beamformer that satisfies the constraint set:

$$\mathbf{H}_0^{\mathrm{H}}\widetilde{\mathbf{G}}=\mathbf{I}. \tag{43}$$

It can be verified that the following definition of $\mathbf{H}_0$

$$\mathbf{H}_0=\widetilde{\mathbf{G}}\left(\widetilde{\mathbf{G}}^{\mathrm{H}}\widetilde{\mathbf{G}}\right)^{-1}, \tag{44}$$

is a proper fixed BF.[1] The matrix $\mathbf{B}$, usually referred to as a blocking matrix, is an $N\times(N-J)$ matrix orthogonal to steering vectors of the speakers such that

$$\mathbf{B}^{\mathrm{H}}\widetilde{\mathbf{G}}=\mathbf{0}. \tag{45}$$

Discussion about various methods for constructing a sparse blocking matrix can be found in [30].

The filter matrix $\mathbf{H}_{\mathrm{NC}}$ is the noise canceller that is responsible for mitigating the residual noise at the outputs of $\mathbf{H}_0$. The closed-form solution for the noise canceller is obtained by minimizing the total noise power at the outputs and is given by

$$\begin{aligned}\mathbf{H}_{\mathrm{NC}}&=\underset{\mathbf{H}}{\operatorname{argmin}}\,\mathrm{Tr}\left[\left(\mathbf{H}_0-\mathbf{B}\mathbf{H}\right)^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}\left(\mathbf{H}_0-\mathbf{B}\mathbf{H}\right)\right]\\&=\left[\mathbf{B}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{B}\right]^{-1}\mathbf{B}\mathbf{\Phi}_{\mathbf{v}}\mathbf{H}_0.\end{aligned} \tag{46}$$

Note that, since $\mathbf{B}$ is usually designed as an $N\times(N-J)$ matrix, $\mathbf{B}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{B}$ is an $(N-J)\times(N-J)$ matrix.

[1] According to our experience, it is recommended to ensure the invertibility of $\widetilde{\mathbf{G}}^{\mathrm{H}}\widetilde{\mathbf{G}}$ by applying regularization to $\widetilde{\mathbf{G}}^{\mathrm{H}}\widetilde{\mathbf{G}}$ in the brackets of (44), yielding $\mathbf{H}_0=\widetilde{\mathbf{G}}\left(\widetilde{\mathbf{G}}^{\mathrm{H}}\widetilde{\mathbf{G}}+\varepsilon\mathbf{I}\right)^{-1}$. In our experiments, we set $\varepsilon=10^{-4}\lambda_{\max}$, where $\lambda_{\max}$ denotes the maximum eigenvalue of $\widetilde{\mathbf{G}}^{\mathrm{H}}\widetilde{\mathbf{G}}$. This way, the condition number of the inverted matrix is constrained to be lower than 10^4.

B. Decision-Directed Based Estimation of $\mathbf{\Phi}_{\mathbf{s}_F}$

Assuming statistical independence between the speakers, $\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}$ can be modelled as a diagonal matrix. To maintain this structure, only the diagonal elements of $\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}$ are estimated and the off-diagonal elements are substituted by zero elements.

In this paper, we adopt the decision-directed approach proposed in [16] (see also [31] and [17]). According to this approach, an estimation of $\phi_{S_{\mathrm{F},j}}$ is obtained by weighting estimates of $\phi_{S_{\mathrm{F},j}}$ from the previous and the current frames, i.e.

$$\begin{aligned}\widehat{\phi}_{S_{\mathrm{F},j}}&=\beta_r\left|\widehat{S}_{\mathrm{LCMV+MCWPF},j}(\ell-1)\right|^2\\&\quad+(1-\beta_r)\max\left\{\left|\widehat{S}_{\mathrm{LCMV},j}(\ell)\right|^2-\phi_{V,\mathrm{RE},j},0\right\},\end{aligned} \tag{47}$$

where $0\le\beta_r<1$ is a weighting factor and the PSD $\phi_{V,\mathrm{RE},j}$ is the jth element of the diagonal of $\mathbf{\Phi}_{\mathbf{v},\mathrm{RE}}$. For the LSA estimator, $\widehat{\mathbf{s}}_{\mathrm{LCMV+MCWPF}}$ may be substituted by $\widehat{\mathbf{s}}_{\mathrm{LCMV+MCLSA}}$,

C. Calculation of $\mathbf{w}_j$ using GSC

The calculation of $\mathbf{w}_j$ defined in (29) requires the inversion of the matrix $\mathbf{\Sigma}_{\bar{j}}$ defined in (30). When the power of the speech signals (excluding the jth speaker) is higher than the noise power, it is likely that the matrix $\mathbf{\Sigma}_{\bar{j}}$ has a high condition-number since the elements of the jth column and the jth row of the term $\left(\mathbf{I}-\mathrm{Diag}\left[\mathbf{i}_j\right]\right)\mathbf{\Phi}_{\mathbf{s}_{\mathrm{F}}}\left(\mathbf{I}-\mathrm{Diag}\left[\mathbf{i}_j\right]\right)^{\mathrm{T}}$ are all zeros. Therefore, instead of using (29), it is recommended that $\mathbf{w}_j$ will be calculated using the GSC structure [25].

The filter vector $\mathbf{w}_j$ is actually the MVDR-BF that can be obtained by solving the following optimization criterion

$$\mathbf{w}_j=\underset{\mathbf{w}}{\operatorname{argmin}}\,\mathbf{w}^{\mathrm{H}}\mathbf{\Sigma}_{\bar{j}}\mathbf{w}\ \text{ s.t. }\ \mathbf{w}^{\mathrm{H}}\mathbf{i}_j=1. \tag{48}$$

Following the GSC formulation, $\mathbf{w}_j$ can be written as

$$\mathbf{w}_j=\mathbf{i}_j-\mathbf{D}_{\bar{j}}\mathbf{d}_j, \tag{49}$$

where $\mathbf{i}_j$ is used as the fixed BF that satisfies the constraint (since $\mathbf{i}_j^{\mathrm{H}}\mathbf{i}_j=1$), $\mathbf{D}_{\bar{j}}$ is the BM that blocks the vector $\mathbf{i}_j$ (i.e., $\mathbf{D}_{\bar{j}}^{\mathrm{T}}\mathbf{i}_j=0$) and $\mathbf{d}_j$ is the noise canceller

$$\mathbf{d}_j=\left(\mathbf{D}_{\bar{j}}^{\mathrm{T}}\mathbf{\Sigma}_{\bar{j}}\mathbf{D}_{\bar{j}}\right)^{-1}\mathbf{D}_{\bar{j}}^{\mathrm{T}}\mathbf{\Sigma}_{\bar{j}}\mathbf{i}_j. \tag{50}$$

Note that the multiplication by $\mathbf{D}_{\bar{j}}$ and $\mathbf{D}_{\bar{j}}^{\mathrm{T}}$ in (50) actually deletes the jth column and jth row of $\mathbf{\Sigma}_{\bar{j}}$, and thus $\mathbf{D}_{\bar{j}}^{\mathrm{T}}\mathbf{\Sigma}_{\bar{j}}\mathbf{D}_{\bar{j}}$ is invertible even when the power of the speech signals (excluding the jth speaker) is higher than the noise power. The matrix $\mathbf{D}_{\bar{j}}$ can be set as the identity matrix without the jth column

$$\mathbf{D}_{\bar{j}}=\left[\begin{array}{ccccccc}\mathbf{i}_1&\dots&\mathbf{i}_{j-1}&\mathbf{i}_{j+1}&\dots&\mathbf{i}_J\end{array}\right], \tag{51}$$

since for each $i\neq j$, $\mathbf{i}_i^{\mathrm{T}}\mathbf{i}_j=0$.

VI. Performance Evaluation

In this section, we evaluate the performance of the proposed estimators, and compare them to other classical postfiltering approaches. In Sec.VI-A, the setup of the experiments is elaborated. The performance evaluation is divided into two main parts. First, the separation ability of the multichannel LCMV and the multichannel Wiener postfilter is investigated in Sec. VI-B. Although satisfying the constraint of the LCMV-BF should result in perfect separation of the speakers, some leakage may occur in the presence of estimation errors. In addition, the multichannel WF postfilter may also cause some leakage between the speakers. Secondly, the estimators under test are described in Sec. VI-C, and are evaluated in terms of the segmental output SNR in Sec. VI-D.

A. Setup

The received microphone signals were constructed by convolving clean speech signals and measured room impulse responses (RIRs) recorded in our acoustic lab [32]. The lab dimensions are $[6; 6; 2.4]$ meters and it is equipped with dedicated panels to control the reverberation level. In our setup, the lab reverberation time was set to either $T_{60} = 0.16$ s or $T_{60} = 0.36$ s. We used four speakers positions at a distance of 2 meters in front of an eight-microphone linear array at various angles. The inter-distances between the microphones were $[3,\ 3,\ 3,\ 8,\ 3,\ 3,\ 3]$ cm. Each experiment includes four sentences, each 4–8 seconds long, two by a male and two by a female. The speakers were positioned at $-90°, -45°, 0°$ and $90°$ relative to the array.

The noise signal vector $\mathbf{v}$ was a summation of two components: i) directional stationary noise $\mathbf{v}_{\text{dir}}$ that is computed by convolving a noise signal from the NOISEX-92 database [33] with RIRs for a speaker located at $45°$ relative to the array at a distance of 2 m, and ii) mutually uncorrelated sensor noise $\mathbf{v}_{\text{sen}}$,i.e.,

$$\mathbf{v} = \mathbf{v}_{\text{dir}} + \mathbf{v}_{\text{sen}}. \tag{52}$$

The level of the sensor noise was 10 dB lower than the directional noise i.e.

$$10\log_{10}\frac{\sum_{k,\ell}\|\mathbf{v}_{\text{dir}}(\ell,k)\|^2}{\sum_{k,\ell}\|\mathbf{v}_{\text{sen}}(\ell,k)\|^2} = 10, \tag{53}$$

where $\|\cdot\|$ is the Euclidean norm. The noise signal $\mathbf{v}$ was added to the speech signals with various input signal-to-noise ratio (SNR) levels

$$\text{iSNR} = 10\log_{10}\frac{\sum_{k,\ell,j}\|\mathbf{x}_j(\ell,k)\|^2}{\sum_{k,\ell}\|\mathbf{v}(\ell,k)\|^2}. \tag{54}$$

An illustration of the geometric setup is given in Fig. 2.

The sampling frequency of the speech signals was set to 16 kHz. The frame length of the STFT was 32 ms with 8 ms between successive time frames (i.e., 512 samples per segment with 25% overlap). To mitigate cyclic convolution artifacts, each segment was zero padded with 512 samples (256 samples before the segment and 256 samples after) such that the length of the discrete Fourier transforms equals 1024. The noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}$, which is non-diagonal, was estimated using time-segments in which all speakers are inactive.

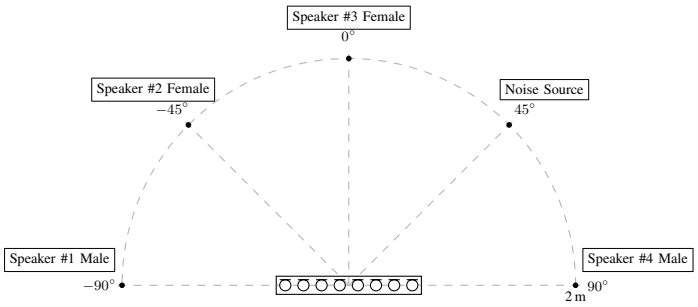


Fig. 2: Geometric setup.

Finally, we set $F_j = G_{1,j}$, indicating that the desired signals are the noiseless replicas of the reverberant speakers at the first microphone, arbitrarily chosen to be the reference microphone. Each RTF $\tilde{\mathbf{g}}_j$ of the various speakers was estimated by the least squares (LS) technique proposed in [25]. For the estimation, time-segments where only one speaker is active were used. To analyze the sensitivity w.r.t. RTF estimation errors due to the additive noise $\mathbf{v}$, we have estimated the RTFs from the noisy signal $\mathbf{y}$, and from the noiseless signal $\tilde{\mathbf{G}}\mathbf{s}$. In this work, a so-called sparse blocking matrix for multiple speakers was used as proposed in [30].

B. Separation Ability

First, the separation ability (i.e., the leakage level) of $\mathbf{H}_{\text{LCMV}}$ and $\mathbf{H}_{\text{WPF}}$ is examined with a representative case where iSNR $= 10$ dB. Only 4×10 signals were tested, i.e., 10 experiments each with 4 concurrent speakers (uttered by 8 microphones). To examine the contributions of the desired speakers alone, the following output signals were evaluated:

1. $\mathbf{u}_{\text{LCMV},j} = \mathbf{H}_{\text{LCMV}}^{\text{H}}\mathbf{x}_j$; (55a)
2. $\mathbf{u}_{\text{WPF},j} = \mathbf{H}_{\text{WPF}}^{\text{H}}\mathbf{i}_j X_{1,j}$; (55b)
3. $\mathbf{u}_{\text{LCMV+MCWPF},j} = \mathbf{H}_{\text{WPF}}^{\text{H}}\mathbf{H}_{\text{LCMV}}^{\text{H}}\mathbf{x}_j$. (55c)

Ideally, for $\mathbf{u}_{\text{LCMV},j}$, we expect to measure the undistorted $X_{1,j}$ at the jth output while all other outputs should be zero (since ideally $\mathbf{u}_{\text{LCMV}} = \mathbf{i}_j X_{1,j}$). Practically, for estimated RTFs, there might be an undesired leakage. In (55b), an oracle LCMV was used to only test the contribution of the MCWF postfilter. In (55c), the outputs of the full multichannel Wiener filter is calculated using the $\mathbf{H}_{\text{LCMV}}$ constructed by estimated RTFs (and thus a leakage is inevitable).

For the purpose of examining the separation ability, we define the following measure, denoted blocking ability ratios (BARs):

$$\text{BAR}_{\text{BF}}(j,i) = \sum_{\ell} 10\log_{10}\frac{\sum_k \left|[\mathbf{u}_{\text{BF},j}]_i\right|^2}{\sum_k |X_{1,j}(\ell,k)|^2}, \tag{56}$$

where, BF $\in \{\text{LCMV}, \text{WF}, \text{LCMV+MCWPF}\}$ and $[\mathbf{u}_{\text{BF},j}]_i$ is the i-th element of vector $\mathbf{u}_{\text{BF},j}$. When $i = j$ BAR measures the distortion of the desired speaker. Ideally, it is expected that $\mathbf{H}_{\text{LCMV}}$ will introduce no attenuation (i.e., $\text{BAR}_{\text{LCMV}}(i,i) = 0$ dB) and full blockage ($\text{BAR}_{\text{LCMV}}(j,i) = -\infty$ dB for $j \neq i$).

First, the results for $\text{BAR}_{\text{LCMV}}(j,i)$ are presented for $T_{60} = 0.36$ in Table I and for $T_{60} = 0.16$ in Table II, where in each

TABLE I: $\text{BAR}_{\text{LCMV}}(j,i)$ for $T_{60} = 0.36$ using RTFs estimated from the noisy signals (top) and using RTFs estimated from noiseless signals (down).

	Output			
Speaker	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	**0.59**	-7.04	-8.33	-7.70
$j=2$	-8.50	**0.37**	-9.43	-7.07
$j=3$	-9.03	-6.63	**1.12**	-6.04
$j=4$	-8.09	-7.55	-7.85	**0.64**

	Output			
Speaker	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	**-0.61**	-11.35	-11.88	-11.22
$j=2$	-10.92	**-0.61**	-12.43	-9.88
$j=3$	-11.90	-10.40	**-0.48**	-9.88
$j=4$	-9.26	-10.85	-10.64	**-0.02**

table the RTFs were estimated using: i) signals with a noise level of iSNR $= 10$ dB, and ii) noiseless signals. In both cases we use time-segments where only the speaker of interest is active. For $i = j$, the results are depicted in boldface. It can be verified that $\text{BAR}_{\text{LCMV}}(j,i)$ elements are close to 0 dB for $i = j$ (diagonal elements). As for $i \neq j$ (off diagonal elements), it can be generally verified that using RTFs estimated from noiseless signals, the blocking ability is better than in the case where the RTFs were estimated from noisy signals. Likewise, the blocking ability for $T_{60} = 0.16$ is better than $T_{60} = 0.36$, since short systems are better described by the RTFs.

Table III depicts the results for $\text{BAR}_{\text{WPF}}(j,i)$ for $T_{60} = 0.36$ and using RTFs estimated from 10 dB noisy signals. The results for $\text{BAR}_{\text{WPF}}(j,i)$ exhibit distortion around -3 dB for $i = j$ and inevitable small leakage, approximately -16 dB, for $i \neq j$. By comparing Tables I and III, it can be concluded that the leakage resulting from the LCMV stage dominates the leakage resulting from the multi-speaker postfiltering stage.

Table IV depicts the results for $\text{BAR}_{\text{LCMV+MCWPF}}(j,i)$ for $T_{60} = 0.36$ and using RTFs estimated from 10 dB noisy signals. Comparing between $\text{BAR}_{\text{LCMV+MCWPF}}(j,i)$ and $\text{BAR}_{\text{LCMV}}(j,i)$ in the same conditions, i.e. Tables I and IV, it can be verified that $\mathbf{H}_{\text{WF}}$ reduces the leakage caused by the LCMV stage. Ideally, the LCMV beamformer should entirely block the interference sources at each output, i.e. each output should be dominated by only one source. In real-life scenarios, due to inevitable estimation errors, leakage of the interference sources is unavoidable. In these cases, $\mathbf{H}_{\text{WF}}$ may further enhance the LCMV beamformer outputs and consequently increase the blocking ability ratio.

C. Noise Reduction: Estimators Under Investigation

In the following sections, we present a comparison of the noise reduction capabilities of various single- and multi-speaker estimators. We evaluated and compared the following five estimators:

1) The multi-speaker LCMV BF:

$$\widehat{\mathbf{s}}_{\text{LCMV}} = \mathbf{H}_{\text{LCMV}}^{\text{H}}\mathbf{y}.$$

The multi-speaker LCMV BF is implemented using the GSC structure, as presented in Sec. V-A.

TABLE II: $\text{BAR}_{\text{LCMV}}(j,i)$ for $T_{60} = 0.16$ using RTFs estimated from the noisy signals (top) and using RTFs estimated from noiseless signals (down).

	Output			
Speaker	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	**0.71**	-10.79	-11.36	-8.36
$j=2$	-9.92	**0.34**	-9.53	-6.90
$j=3$	-9.48	-9.44	**0.88**	-6.90
$j=4$	-10.64	-11.44	-10.69	**0.12**

	Output			
Speaker	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	**-0.72**	-14.86	-14.74	-11.55
$j=2$	-12.28	**-0.37**	-13.00	-7.94
$j=3$	-13.03	-14.61	**-1.13**	-10.58
$j=4$	-11.91	-15.71	-12.66	**-0.59**

TABLE III: $\text{BAR}_{\text{WPF}}(j,i)$ for $T_{60} = 0.36$ and using RTFs estimated from the noisy signals.

	Output			
Speaker	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	**-2.87**	-14.88	-19.35	-17.53
$j=2$	-15.38	**-3.34**	-19.30	-16.60
$j=3$	-17.09	-14.73	**-4.23**	-16.34
$j=4$	-15.23	-16.80	-18.33	**-3.44**

2) The multi-speaker LCMV BF when only the reverberant speech was used as an input (i.e., the noiseless signals) such that the separation performance of the LCMV BF can be examined:

$$\widehat{\mathbf{s}}_{\text{LCMV-NL}} = \mathbf{H}_{\text{LCMV}}^{\text{H}} \sum_{j=1}^{J} \mathbf{x}_j = \mathbf{H}_{\text{LCMV}}^{\text{H}} \mathbf{G}\,\mathbf{s}.$$

Note that the LCMV filters are computed assuming that the noise is present.

3) The multi-speaker LCMV followed by the distortionless spatial filter $\mathbf{Q}$ as defined in (39):

$$\widehat{\mathbf{s}}_{\text{LCMV+Q}} = \mathbf{Q}^{\text{H}}\mathbf{H}_{\text{LCMV}}^{\text{H}}\mathbf{y}.$$

4) The multi-speaker LCMV BF followed by a single-channel Wiener postfilter which is applied to each output:

$$\widehat{\mathbf{s}}_{\text{LCMV+SCWPF}} = \mathbf{H}_{\text{SCWPF}}^{\text{H}}\widehat{\mathbf{s}}_{\text{LCMV}}, \tag{57}$$

where $\mathbf{H}_{\text{SCWPF}} \equiv \text{Diag}\left[\begin{array}{ccc} H_{\text{SCWPF},1} & \dots & H_{\text{SCWPF},J} \end{array}\right]$, $H_{\text{SCWPF},j}$ was evaluated similarly to $H_{\text{WPF},j}$ in (41), with the single-channel a priori SNRs $\xi_j = \frac{\phi_{S_{j,\text{F}}}}{\phi_{V,\text{RE},j}}$.

TABLE IV: $\text{BAR}_{\text{LCMV+MCWPF}}(j,i)$ for $T_{60} = 0.36$ and using RTFs estimated from the noisy signals.

	Output			
Speaker	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	**-2.93**	-14.28	-18.50	-16.49
$j=2$	-14.79	**-3.41**	-18.24	-15.18
$j=3$	-16.31	-13.75	**-4.19**	-15.25
$j=4$	-13.70	-15.47	-16.94	**-3.40**

The PSD $\phi_{S_{\mathrm{F},j}}$ was estimated using the (single-channel) decision-directed approach in (47) where $\widehat{S}_{\mathrm{LCMV+MCWPF},j}$ was replaced by $\widehat{S}_{\mathrm{LCMV+SCWPF},j}$.

5) The multi-speaker LCMV beamformer followed by the multi-speaker Wiener postfilter:

$$\widehat{\mathbf{s}}_{\mathrm{LCMV+MCWPF}} = \mathbf{H}_{\mathrm{WPF}}^{\mathrm{H}} \widehat{\mathbf{s}}_{\mathrm{LCMV}}.$$

The multi-speaker Wiener postfilter $\mathbf{H}_{\mathrm{WPF}}$ was calculated using (40).

6) The LCMV BF with single-channel LSA postfilter:

$$\widehat{\mathbf{s}}_{\mathrm{LCMV+SCLSA}} = \mathbf{H}_{\mathrm{SCLSA}}^{\mathrm{H}} \widehat{\mathbf{s}}_{\mathrm{LCMV}}, \tag{58}$$

where $\mathbf{H}_{\mathrm{SCLSA}} \equiv \mathrm{Diag}\left[\begin{array}{ccc} H_{\mathrm{SCLSA},1} & \cdots & H_{\mathrm{SCLSA},J} \end{array}\right]$, $H_{\mathrm{SCLSA},j}$ was evaluated similarly to $H_{\mathrm{LSA},j}$ in (33), with the single-channel a priori and a posteriori SNRs $\xi_j = \frac{\phi_{S_{j,\mathrm{F}}}}{\phi_{V,\mathrm{RE},j}}$ and $\gamma_j = \frac{\left|\widehat{S}_{\mathrm{LCMV},j}\right|^2}{\phi_{V,\mathrm{RE},j}}$. The PSD $\phi_{S_{\mathrm{F},j}}$ was estimated using the (single-channel) decision-directed approach in (47) where $\widehat{S}_{\mathrm{LCMV+MCWPF},j}$ was replaced by $\widehat{S}_{\mathrm{LCMV+SCLSA},j}$.

7) The LCMV BF with the multi-speaker LSA estimator from (37):

$$\widehat{\mathbf{s}}_{\mathrm{LCMV+MCLSA}} = \mathbf{H}_{\mathrm{LSA}}^{\mathrm{H}} \widehat{\mathbf{s}}_{\mathrm{LCMV}}. \tag{59}$$

D. Interference and Noise Reduction Results

The aforementioned estimators were compared in terms of output segmental signal-to-interference plus noise ratio (SINR). The output segmental SINR was aggregated for all speakers and for all time-frequency bins and is defined as

$$\mathrm{oSINR}_{\mathrm{BF}} = \sum_{\ell} 10\log_{10} \frac{\sum_k \|\mathbf{s}_{\mathrm{F}}(\ell,k)\|^2}{\sum_k \|\widehat{\mathbf{s}}_{\mathrm{BF}}(\ell,k) - \mathbf{s}_{\mathrm{F}}(\ell,k)\|^2}, \tag{60}$$

where

$$\mathrm{BF} \in \big\{\mathrm{LCMV}, \mathrm{LCMV+Q}, \mathrm{LCMV+SCWPF}, \mathrm{LCMV+MCWPF}, \mathrm{LCMV+SCLSA}, \mathrm{LCMV+MCLSA}\big\}.$$

In addition, the input SINR was calculated, i.e., the SINR in (60) where $\widehat{\mathbf{s}}_{\mathrm{BF}}(\ell,k)$ is substituted by $Y_1(\ell,k) \times \left[\begin{array}{cccc} 1 & 1 & 1 & 1 \end{array}\right]^{\mathrm{T}}$. All measurements were computed by averaging the output segmental SINR results obtained using 4×50 sentences, i.e., 50 experiments for each scenario where each experiment consists of 4 concurrent speakers (uttered by 8 microphones). The weighting factor β_r was set to 0.99. The lower-bound gain $H_{\min}$ was set to 0.1.

In Tables V and VI, the output segmental SINR results are presented for $T_{60} = 0.16$ sec and $T_{60} = 0.36$ sec. The best results are depicted in boldface font. The separation ability of the LCMV BF can be examined from the results of the LCMV BF that is applied to the noiseless signals. Generally, an improvement of 11 dB is obtained compared to the oSINR of the unprocessed signal. The performance of the standard LCMV BF depends on the input SNR and is generally lower than the performance of the LCMV BFs with the postfilter. Additionally, the performance of LCMV+Q is generally higher than the performance of the standard LCMV BF but lower than the performance of the LCMV BFs with the postfilter. It can be verified that the proposed multi-speaker algorithms (LCMV+MCWPF and LCMV+MCLSA) outperform their single-speaker counterparts (namely, LCMV+SCWPF and LCMV+SCLSA).

TABLE V: oSINR for the various estimators ($T_{60} = 0.16$ s)

	iSNR			
Method	0 dB	5 dB	10 dB	15 dB
Unprocessed	-13.64	-10.86	-9.03	-7.87
LCMV-NL	-1.77	0.18	3.45	5.16
LCMV	-9.95	-7.16	-1.96	1.64
LCMV+Q	-4.188	-2.18	0.80	2.86
LCMV+SCWPF	-1.64	0.68	3.61	4.87
LCMV+MCWPF	**0.48**	**2.91**	**4.85**	**5.48**
LCMV+SCLSA	-1.77	0.54	3.49	4.79
LCMV+MCLSA	0.11	2.80	4.80	5.42

TABLE VI: oSINR for the various estimators ($T_{60} = 0.36$ s)

	iSNR			
Method	0 dB	5 dB	10 dB	15 dB
Unprocessed	-13.78	-11.00	-9.16	-7.98
LCMV-NL	-1.90	0.00	2.78	4.39
LCMV	-11.89	-8.53	-3.29	0.42
LCMV+Q	-4.04	-2.48	0.05	2.00
LCMV+SCWPF	-1.68	0.55	3.05	4.24
LCMV+MCWPF	**0.52**	**2.94**	**4.64**	**5.18**
LCMV+SCLSA	-1.80	0.43	2.93	4.13
LCMV+MCLSA	-0.03	2.69	4.55	5.12

Example sonograms of the various output signals for input SNR of 15 dB and $T_{60} = 0.36$ s are depicted in Fig. 3. Figure 3a depicts $X_{1,1}$, the reverberant signal of speaker 1 (positioned at $-90°$), as received by the reference microphone. Figure 3b depicts $\sum_j X_{1,j}$, the reverberant signals of all speakers as concurrently received by the reference microphone, while Fig. 3c depicts Y_1, the total received signal (including the noise). Figure 3d depicts $\widehat{S}_{\mathrm{LC},1}$, the first component of the multi-speaker LCMV output, corresponding to the first speaker. Likewise, Figures 3e, 3f, 3g and 3h depict the outputs $\widehat{S}_{\mathrm{LCMV+SCWPF},1}$, $\widehat{S}_{\mathrm{LCMV+MCWPF},1}$, $\widehat{S}_{\mathrm{LCMV+SCLSA},1}$ and $\widehat{S}_{\mathrm{LCMV+MCLSA},1}$, respectively.

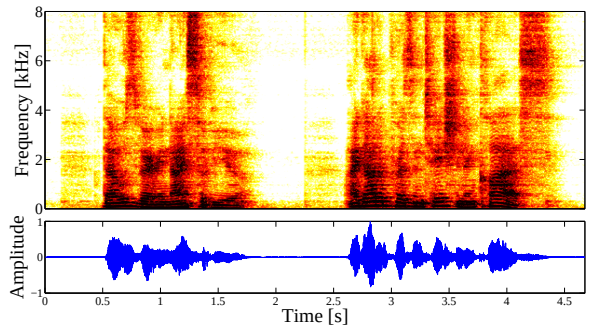
The separation ability of the multi-speaker LCMV can be seen in Fig. 3d. However, the LCMV exhibits poor noise reduction. The postfilter outputs, i.e., Figures. 3e, 3f, 3g and 3h, exhibit better noise reduction than the LCMV output. Beyond that, using careful examination, it can be seen that the multi-speaker postfilters reduce more noise than the single-speaker postfilters (i.e., Fig. 3f relative to Fig. 3e and Fig. 3h relative to Fig. 3g).

Audio examples are available in our website.[2] By listening to these examples, it is evident that the proposed multi-speaker estimators produces the highest noise reduction, when compared with the baseline single-speaker estimators.

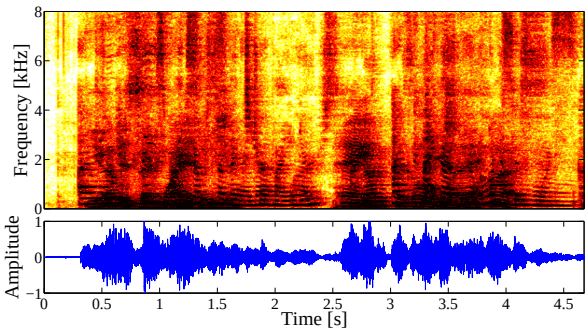
VII. Conclusions

In the current paper, the MMSE estimator of several concurrent speech signals in noisy environment was decomposed into multichannel LCMV beamformer followed by multichannel

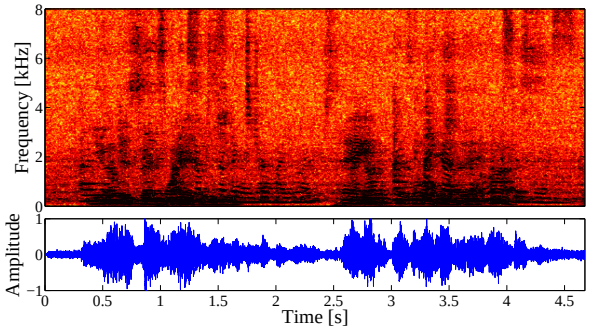
[2] http://www.eng.biu.ac.il/gannot/speech-enhancement/



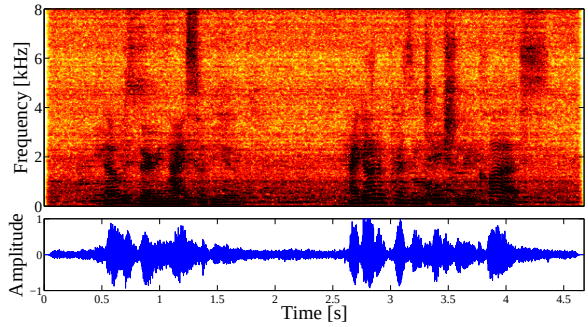
(a) Reverberant speech of speaker #1.



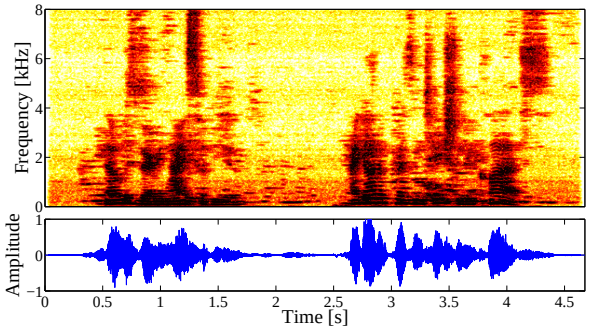
(b) Reverberant speech of all concurrent speakers.



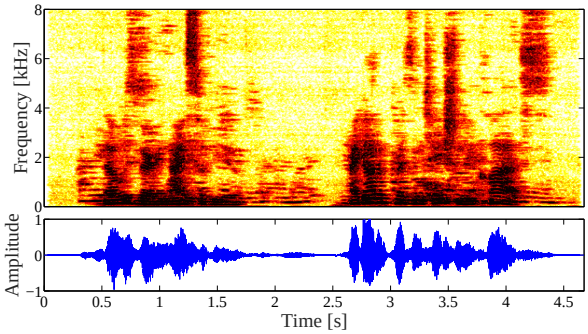
(c) Input signal of microphone #1.



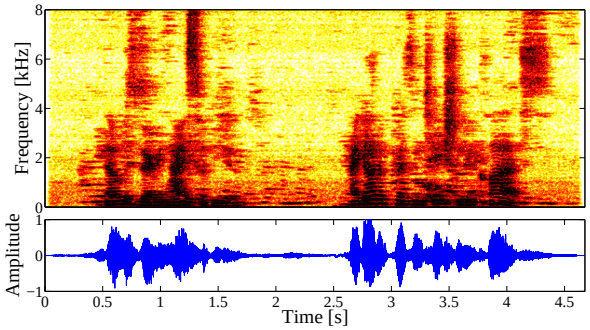
(d) LCMV output.



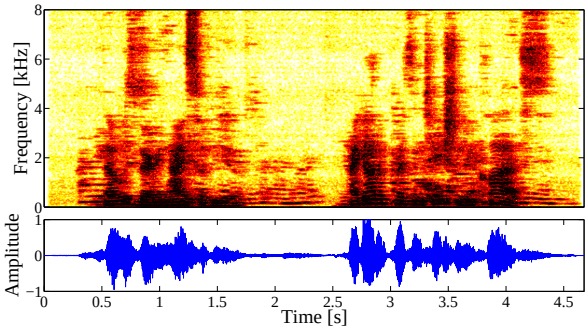
(e) LCMV followed by single-channel Wiener filter.



(f) LCMV followed by the multispeaker Wiener postfilter.



(g) LCMV followed by single-channel LSA.



(h) LCMV followed by multi-speaker LSA.

Fig. 3: Spectrograms with $T_{60} = 0.36$ s and input SNR of 15 dB.

Wiener postfilter. The output of the multichannel LCMV beamformer was proved to be the sufficient statistic of the measurements for estimating the speech signals. Also the multi-speaker LSA estimator was derived. The algorithms were tested in a room with a reverberation time of 0.16 s and 0.36 s for several signal-to-noise levels of directional noise. In terms of output SNR the proposed multi-speaker algorithms significantly outperform competing single-channel postfiltering algorithms.

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Proc. Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[2] M. H. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Transactions Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1378–1393, 1983.

[3] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 168–169, 2002.

[4] S. Werner, J. Apolinário, M. L. De Campos *et al.*, "On the equivalence of RLS implementations of LCMV and GSC processors," *IEEE Signal Process. Lett.*, vol. 10, no. 12, pp. 356–359, 2003.

[5] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[6] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Transactions Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[7] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[8] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. Ray Liu, Eds. Wiley, 2010, ch. 9.

[9] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 3, pp. 39–60.

[10] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New-York, USA, Apr. 1988, pp. 2578–2581.

[11] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[12] S. Leukimmiatis, D. Dimitriadis, and P. Maragos, "An optimum microphone array post-filter for speech applications," in *Proc. Interspeech Conf.*, 2006, pp. 2142–2145.

[13] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. 1–4.

[14] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1064–1073, 2003.

[15] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

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

[18] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2002, pp. 209–213.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions Audio, Speech, Lang. Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[20] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.

[21] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[22] O. Thiergart, M. Taseska, and E. A. Habets, "An informed MMSE filter based on multiple instantaneous direction-of-arrival estimates," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[23] H. L. Van Trees, *Detection, estimation, and modulation theory, optimum array processing*. John Wiley & Sons, 2004.

[24] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley, Reading, MA, 1991.

[25] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[26] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Tech. Rep., 2012.

[27] M. J. Schervish, *Theory of statistics*. Springer, 1995.

[28] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1979, pp. 208–211.

[29] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[30] S. Markovich-Golan, S. Gannot, and I. Cohen, "A sparse blocking matrix for multiple constraints GSC beamformer," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 197–200.

[31] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, 2004, pp. 250–254.

[32] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

[33] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.