Successive relative transfer function identification using blind oblique projection

Dani Cherkassky, *Student member, IEEE,* and Sharon Gannot, *Senior member, IEEE*

Distortionless speech extraction in a reverberant environment can be achieved by applying a beamforming algorithm, provided that the relative transfer functions (RTFs) of the sources and the covariance matrix of the noise are known. In this paper, the challenge of RTF identification in a multi-speaker scenario is addressed. We propose a successive RTF identification (SRI) technique, based on the sole assumption that sources do not become simultaneously active. That is, we address the challenge of estimating the RTF of a specific speech source while assuming that the RTFs of all other active sources in the environment were previously estimated in an earlier stage. The RTF of interest is identified by applying the blind oblique projection (BOP)-SRI technique. When a new speech source is identified, the BOP algorithm is applied. BOP results in a null steering toward the RTF of interest, by means of applying an oblique projection to the microphone measurements. We prove that by artificially increasing the rank of the range of the projection matrix, the RTF of interest can be identified. An experimental study is carried out to evaluate the performance of the BOP-SRI algorithm in various signal to noise ratio (SNR) and signal to interference ratio (SIR) conditions and to demonstrate its effectiveness in speech extraction tasks.

***Index Terms*—Relative transfer function, system identification, oblique projection.**

I. Introduction

SPEECH enhancement and separation are fundamental challenges in audio signals processing. In many speech processing applications, such as hands-free telephony, human-machine interface and hearing aids, the received signal is a mixture of the desired speech, one or more interfering sources, such as competing speakers, and background noise. The presence of the interfering sources cause a signal degradation which can lead to an unintelligibility of the speech and to severe degradation in speech recognition systems performance. Hundreds of multichannel speech enhancement techniques have been proposed in the literature over the last decades. Well known techniques include: non-negative matrix factorization [1], [2], blind source separation [3], [4], and beamforming [5], [6]. Recently, deep neural networks are also suggested to address the speech extraction challenge, [7], [8]. A comprehensive literature review and comparison of the aforementioned techniques is given in [9]. The current work is focused on a speech extraction/separation challenge in a reverberant environment, by an application of a beamformer.

The relative transfer function (RTF) is an important component of multi-microphone speech processing systems, particularly in reverberant environments [10]–[15]. An RTF describes the coupling between the signals received at the microphones as a response to a single source. One of the most common applications of RTFs is speech extraction in a noisy and reverberant environment. For instance, the constraint set of the linearly constrained minimum variance (LCMV) beamformer can be expressed in terms of the sources' RTFs [16]. A formulation of the constraints set in terms of RTFs allows the LCMV beamformer to reject the interfering speech without distorting the desired speech components.

The RTF identification challenge in a noisy environment with a single active speech source has been well studied. Shalvi and Weinstein [17] proposed identifying the coupling between speech components received at two microphones by using the nonstationarity of the desired speech signal received at the sensors, assuming stationary additive noise and time-invariant RTF. The observed signal is divided into subintervals. The speech signal is regarded as stationary in each subinterval and nonstationary between subintervals. Accordingly, an overdetermined set of equations for two unknown variables, the RTF and the cross power spectral density (PSD) of the noise signals, can be formulated by computing the PSD of the sensor signals in each subinterval. The estimates of these two variables are derived by applying the weighted least squares approach. Cohen [18] proposed an RTF identification method that utilizes the speech presence probability (SPP) in the time-frequency domain to identify the time-frequency instances that contain the speech signal. By using the SPP, it is possible to cluster the subintervals into two groups, one consisting of noise-only subintervals and the second of subintervals in which speech is present. The first group is utilized for estimating the noise cross PSD, while the second group is utilized to derive an RTF estimator. More recently, two RTF identification methods were the topic of an intensive study: the covariance subtraction (CS) method [19], [20] and the covariance whitening (CW) [16], [21] method. Both methods assume that the information on the activity pattern of the speech sources of interest is available and utilize the signals' PSD matrices, obtained during noise-only time-segments and during speech plus noise time-segments, to estimate the RTF. A comparative survey of the CS and CW methods for RTF estimation was presented by Markovich-Golan and Gannot [22].

RTF estimation in a multiple and concurrent speaker scenario was recently considered in the literature. Markovich-Golan et al. [23] proposed tracking the desired and interference speakers' subspaces in non-static scenarios with concurrently active multiple speakers in a reverberant environment. It was proven by Hadad et al. [24] that knowledge of a basis that spans the subspace of the desired sources and a basis that spans the subspace of the interfering sources suffices for

D. Cherkassky and S. Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel (e-mail: dani.cherkassky@gmail.com; sharon.gannot@biu.ac.il).

implementing the LCMV beamforming algorithm. The aforementioned desired and interfering sources' subspaces can be estimated in a scenario where both all the desired sources and all the interfering sources are simultaneously active. However, signal segments in which both the desired and interfering sources are concurrently active cannot be used for estimating the subspaces. Hassani et al. [25] proposed a method for estimating the desired and the interfering sources' subspaces by exploiting signal segments having concurrent activity of the desired and the interfering sources. It was assumed that an initial estimate of the desired and interfering sources' subspaces is available, and then, the individual subspace estimates were projected onto the joint signal subspace of all the desired and interfering sources. The procedure exploits signal segments with concurrent activity of the desired and the interfering sources and results in an improved estimate of the individual subspaces as compared with the initially available estimates. Deleforge et al. [26] proposed a generalization of the RTFs' definition to several sources. The generalized RTFs are defined through multichannel, multi-frame spectrograms of the received noise-free signal. Markovich-Golan et al. [27] suggested using the Triple-N ICA for convolutive mixtures (TRINICON) [28], a blind source separation (BSS) framework for estimating RTFs in a multi-speakers scenario. The proposed algorithm assumes the availability of an initial, direct-path based estimate of the target RTFs [29].

In this paper, we consider a multi-source scenario. We propose a successive RTF identification (SRI) technique based on the sole assumption that sources do not become simultaneously active. Namely, we address the challenge of estimating the RTF of a specific speech source while assuming that the RTFs of all other active sources in the environment were previously estimated. The proposed SRI algorithm is founded on an oblique (nonorthogonal) projection operator [30]. Oblique projection is used to project measurements into a low-rank subspace along a direction that is oblique to the subspace. Unlike orthogonal projection, oblique projection provides the flexibility to design a nonorthogonal null space and range. In this contribution, we introduce the blind oblique projection (BOP) for RTF estimation. The range of the BOP is set to include all the previously estimated RTFs, while the null space of the BOP is designed blindly to include the RTF of interest. Specifically, we propose resolving the null space of the BOP by minimizing the norm of the projected measurements, subject to keeping the range of the BOP fixed. The above described RTF estimation procedure is referred to in the following as BOP-SRI. In order to demonstrate that the BOP-SRI provides a valid RTF estimator, we prove that, if the dimension of the range of the BOP is equal to the number of microphones minus one and assuming a sufficiently high signal to noise ratio (SNR), the resulting null space of the BOP will be set to a vector parallel to the RTF of interest. It should be noted that in the case where the number of active sources in the environment is lower than the number of microphones, which is the case in typical speech enchantment applications, the dimension of the range can be artificially increased to facilitate BOP-SRI implementation. At this point, it may be worth stressing that the proposed technique was evaluated in controlled environments, both simulated and actual acoustic lab. Applying the proposed technique to a more complex scenarios with arbitrary activity pattern of the speakers, may require more sophisticated speakers counting and RTF association algorithms.

The rest of the paper is organized as follows. Section II is dedicated to the introduction of the considered signal model and the properties of the oblique projection operator. In Section III, we derive the BOP-SRI algorithm and present some practical considerations that should be addressed when utilizing BOP-SRI in a typical speech extraction application. An experimental study, evaluating the performance of the proposed BOP-SRI algorithm, is presented in Section V. Finally, we conclude the work with a brief discussion in Section VI.

II. Preliminaries

This section is split into two parts. In the first part, we present the considered signals and measurement model and, in the second part, the oblique projection operator. Oblique projection is a well established mathematical tool. However, it has received relatively little attention in the multi-microphone speech processing literature. Accordingly, since the oblique projection operator is the core of the proposed BOP-SRI algorithm, hereafter we present the main properties of the operator.

A. Signal model

Consider an array consisting of M microphones capturing a time-varying acoustical scene. Each of the involved signals propagates through the acoustic environment before being picked up by the array. In the short-time Fourier transform (STFT) domain, the nth speech source is denoted by $s_n(\ell,k)$, the acoustic transfer function (ATF) relating the nth source and the mth microphone by $g_{m,n}(k)$, and the stationary noise at the mth microphone by $v_m(\ell,k)$, where ℓ is the frame index and k is the frequency index. The received signals in the STFT domain can be formulated in a vector representation:

$$\mathbf{z}(\ell,k)=\sum_{n=1}^{N}\mathcal{I}_n(\ell)\mathbf{h}_n(k)x_n(\ell,k)+\mathbf{v}(\ell,k), \tag{1}$$

where N is the number of sources of interest, $x_n(\ell,k)=g_{1,n}(k)s_n(\ell,k)$, $\mathcal{I}_n(\ell)\in\{0,1\}$ indicates the activity of $s_n(\ell,k)$, and $\mathbf{h}_n(k)$ is the RTF vector of the nth source defined as

$$\mathbf{h}_n(k)=\left[1,\frac{g_{2,n}(k)}{g_{1,n}(k)},\cdots,\frac{g_{M,n}(k)}{g_{1,n}(k)}\right]^T. \tag{2}$$

Considering the sources' activity pattern, we assume that the speech sources become active successively. Accordingly, the activity indicator function of the nth source is defined by

$$\mathcal{I}_n(\ell)=\begin{cases}0, & \text{if } \ell\le\ell_n\\ 1, & \text{if } \ell_n<\ell\le\ell_{n+1}\\ \mathcal{A}_n, & \text{otherwise.}\end{cases} \tag{3}$$

where $\mathcal{A}_n \in \{0,1\}$. The noise $\mathbf{v}(\ell,k)$ is assumed active throughout the measurement period. The considered activity pattern may be practical, for example, in a noisy conference call scenario. In such a scenario, typically, the speech sources do not become simultaneously active but do remain active for a sufficient amount of time before they become inactive again. Accordingly, the proposed $\mathcal{I}_n(\ell)$ dictates a unique activation time ℓ_n of the nth speech source. Upon activation, the nth source remains active for at least $\ell_n < \ell \leq \ell_{n+1}$ time frames, while for time frames $\ell > \ell_{n+1}$ the nth source can be either active or inactive, as suggested by the definition of $\mathcal{A}_n$. However, we do assume that simultaneous activation and deactivation of two independent sources never occur. The probability of simultaneous activation and deactivation of two independent sources was inquired in [31], and it was suggested that the probability of such an event is zero. In practice, the activity pattern of the sources is, of course, unknown and should be estimated from the measurements. Source activity function estimation is addressed in Section IV. It should be stressed that the BOP-SRI algorithm, proposed in the sequel, addresses time frames where multiple speech sources are simultaneously active in order to estimate the target RTF.

B. Oblique projection operator

In signal processing applications, oblique projections are used to project measurements onto a low-rank subspace along a direction that is oblique to the subspace. The SRI algorithm presented in Section III is based on the oblique projection operator, and therefore we review the main properties of this operator in this section.

Consider an M-dimensional measurement vector $\mathbf{z} = \mathbf{s}+\mathbf{n}$, where the signal $\mathbf{s} = \mathbf{H}\mathbf{x}$ lies in an $N-1$ dimensional subspace of $\mathbb{C}^M$, which we denote by $\langle \mathbf{H} \rangle$. The subspace $\langle \mathbf{H} \rangle$ is the range of the transformation $\mathbf{H}$ and is spanned by the columns of the matrix $\mathbf{H}$. These columns comprise a basis for the subspace, and the elements of $\mathbf{x} = [x_1, x_2, \cdots, x_{N-1}]^T$ are the coordinates of $\mathbf{w}$ with respect to this basis. Similarly, $\mathbf{n} = \mathbf{h}x_N$ lies in a 1-dimensional subspace of $\mathbb{C}^M$. This subspace, spanned by the vector $\mathbf{h}$, is denoted by $\langle \mathbf{h} \rangle$.

An oblique projection $\mathbf{E_{Hh}}$, of which the range is $\langle \mathbf{H} \rangle$ and the null space comprises $\langle \mathbf{h} \rangle$, is defined by [32]:

$$\mathbf{E_{Hh}} = \mathbf{H}\left(\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}, \tag{4}$$

where $\mathbf{P}_{\mathbf{h}}^{\perp} = \mathbf{I} - \mathbf{h}\left(\mathbf{h}^H\mathbf{h}\right)^{-1}\mathbf{h}^H$ is an orthogonal projection matrix to the null subspace of $\langle \mathbf{h} \rangle$ and $\mathbf{I}$ is an identity matrix. Equivalently, $\mathbf{E_{hH}}$ is an oblique projection with range $\langle \mathbf{h} \rangle$ and with the null space comprising $\langle \mathbf{H} \rangle$. It is straightforward to verify that $\mathbf{E_{Hh}}$ is an idempotent with range $\langle \mathbf{H} \rangle$ and a null space that includes $\langle \mathbf{h} \rangle$

$$\mathbf{E_{Hh}}\mathbf{H} = \mathbf{H}\left(\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H} = \mathbf{H}, \tag{5a}$$

$$\mathbf{E_{Hh}}\mathbf{h} = \mathbf{H}\left(\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{h} = \mathbf{0}. \tag{5b}$$

To complete the null space, let us define a matrix $\mathbf{A}$, the columns of which span the $M-N$ dimensional subspace perpendicular to $\langle \mathbf{H}\ \mathbf{h} \rangle$. By definition, $\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{A} = \mathbf{A}$ and $\mathbf{H}^H\mathbf{A} = \mathbf{0}$, accordingly

$$\begin{aligned}\mathbf{E_{Hh}}\mathbf{A} &= \mathbf{H}\left(\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{A} = \\ &= \mathbf{H}\left(\mathbf{H}^H\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{A} = \mathbf{0}.\end{aligned} \tag{6}$$

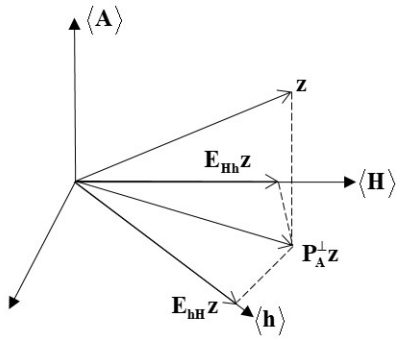

Fig. 1: Geometrical interpretation of the oblique projection in the Euclidean space.

In Fig. 1, we show the geometrical interpretation of the oblique projection operator in the Euclidean space. As shown, $\langle \mathbf{A} \rangle$ is the subspace orthogonal to both $\langle \mathbf{H} \rangle$ and $\langle \mathbf{h} \rangle$ and $\mathbf{E_{Hh}}$ is a projection operator with a range equal to $\langle \mathbf{H} \rangle$ and a null space equal to $\langle \mathbf{h}\ \mathbf{A} \rangle$. In a special case, where the unification of $\langle \mathbf{H} \rangle$ and $\langle \mathbf{h} \rangle$ spans the entire Euclidean space, i.e., $\langle \mathbf{A} \rangle$ is an empty subspace, the null space of $\mathbf{E_{Hh}}$ is equal to $\langle \mathbf{h} \rangle$. It is noteworthy to mention that the ability of the oblique projection operator to project onto $\langle \mathbf{H} \rangle$ while nulling the subspace $\langle \mathbf{h} \rangle$ that is nonorthogonal to $\langle \mathbf{H} \rangle$ comes with a price. Applying $\mathbf{E_{Hh}}$ to a random vector $\mathbf{v}$ may result in an increase in $||\mathbf{E_{Hh}}\mathbf{v}||_2$ as compared to $||\mathbf{v}||_2$ [32]. The level of the increase in the norm depends on the principal angle [33] between subspaces $\langle \mathbf{H} \rangle$ and $\langle \mathbf{h} \rangle$: the smaller the principal angle, the higher is the expected amplification [34]. A similar phenomenon of noise increase by an LCMV beamformer in a scenario where the desired and interfering speakers are spatially close was demonstrated in [35].

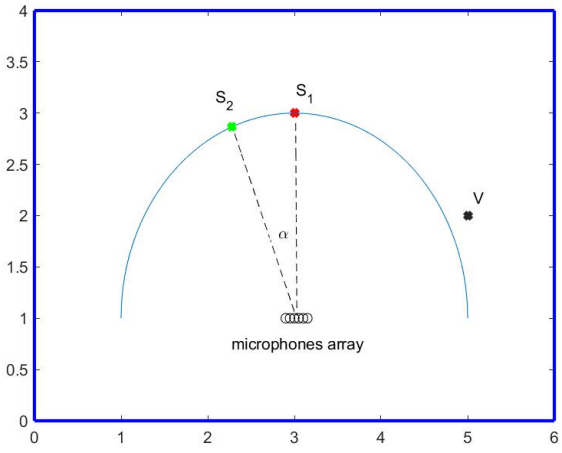

Fig. 2: Simulative setup.

Let us quantify the noise amplification phenomenon by the

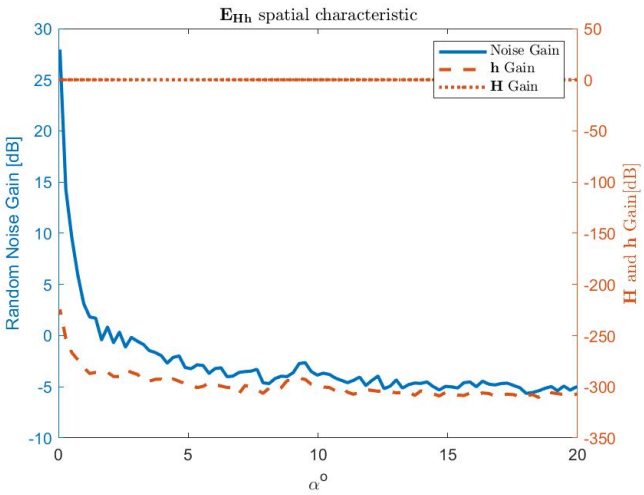

Fig. 3: $\mathbf{E_{Hh}}$ Gains towards the sources of interest.

following simulative study. We simulated a microphone array with 8 microphones placed in a $6 \times 4 \times 2.5$ m room with $T_{60} \approx 500$ mSec. A top view of the room is depicted in Fig. 2. Two speech sources, s_1 and s_2, with RTFs denoted by $\mathbf{H}$ and $\mathbf{h}$, respectively, are located on an arc with a radius of 1.5 m, centered on a microphone array center. A noise source v is also present in the environment. In Fig. 3 we present the responses of $\mathbf{E_{Hh}}$ towards the considered RTFs, as well as towards the noise source, as a function of α, the angle between s_1 and s_2 locations. The simulative results demonstrate that applying $\mathbf{E_{Hh}}$ will amplify the random noise in the environment when principal angle between $\mathbf{H}$ and $\mathbf{h}$ is small. In the simulated scenario noise amplification is observed for $\alpha < 3^{\circ}$.

In the following section, we utilize the oblique projection operator for deriving the proposed BOP-SRI algorithm.

III. Successive Relative Transfer function identification algorithm

When multiple speech sources are concurrently active, the RTF identification techniques that assume a single speech source in a noisy environment are not valid. In the following, we propose the BOP-SRI algorithm for $\mathbf{h}_{n_0}(k)$ identification, under the assumption that the RTFs $\mathbf{h}_n(k), n < n_0$ of all already active sources in the environment were previously identified. A similar challenge was considered in [36], where a single microphone speech enhancement technique was utilized for identifying $\mathbf{h}_{n_0}(k)$. Hereafter, we address the SRI challenge by applying the BOP technique.

The idea behind the BOP algorithm is to set the range of the projection to the subspace spanned by the previously identified RTFs $\mathbf{h}_n(k), n < n_0$, followed by the optimization of the null space such that the power of the projected measurements is minimal. In the following, we demonstrate that minimal power is achieved by nulling out the signal from $\mathbf{h}_{n_0}(k)$. We also prove that, when $\langle \mathbf{A} \rangle$ is an empty subspace, nulling out the signal from $\mathbf{h}_{n_0}(k)$ is equivalent to $\mathbf{h}_{n_0}(k)$ identification.

Assuming the past speech signals remain active, $\mathcal{A}_n = 1, n = 1, \ldots, n_0 - 1$, the received signal (1) in frames $\ell_{n_0} < \ell < \ell_{n_0+1}$ can be formulated in a matrix notation:

$$\mathbf{z}(\ell,k) = \mathbf{H}(k)\mathbf{x}(\ell,k) + \mathbf{h}_{n_0}(k)x_{n_0}(\ell,k) + \mathbf{v}(\ell,k), \quad (7)$$

where $\mathbf{x}(\ell,k) = [x_1(\ell,k), x_2(\ell,k), \cdots, x_{n_0-1}(\ell,k)]^T$, $\mathbf{H}(k) = [\mathbf{h}_1(k), \mathbf{h}_2(k), \cdots, \mathbf{h}_{n_0-1}(k)]$ and $\mathbf{v}(\ell,k)$ is the noise. For simplicity, the frequency index is omitted hereinafter.

Let us show that $\mathbf{E}_{\mathbf{Hh}_{n_0}}$ minimizes the power of the projected measurements, under a plausible assumption of a sufficiently large SNR. Applying $\mathbf{E}_{\mathbf{Hh}_{n_0}}$ to the received signal results in

$$\mathbf{y}(\ell;\mathbf{h}_{n_0}) = \mathbf{E}_{\mathbf{Hh}_{n_0}}\mathbf{z}(\ell) \approx \mathbf{H}\mathbf{x}(\ell) + \mathbf{E}_{\mathbf{Hh}_{n_0}}\mathbf{v}(\ell). \quad (8)$$

Since the sources $s_n, 0 < n \leq n_0$ are mutually uncorrelated and all are uncorrelated with the noise $\mathbf{v}$, the power of $\mathbf{z}(\ell)$ and $\mathbf{y}(\ell;\mathbf{h}_{n_0})$ is given by

$$\begin{aligned} E\left\{\mathbf{z}^H(\ell)\mathbf{z}(\ell)\right\} &= E\left\{\mathbf{x}^H(\ell)\mathbf{H}^H\mathbf{H}\mathbf{x}(\ell)\right\} + \\ &+ E\left\{x^*_{n_0}(\ell)\mathbf{h}^H_{n_0}\mathbf{h}_{n_0}x_{n_0}(\ell)\right\} + \\ &+ E\left\{\mathbf{v}^H(\ell)\mathbf{v}(\ell)\right\}, \quad (9a) \\ E\left\{\mathbf{y}^H(\ell;\mathbf{h}_{n_0})\mathbf{y}(\ell;\mathbf{h}_{n_0})\right\} &= E\left\{\mathbf{x}^H(\ell)\mathbf{H}^H\mathbf{H}\mathbf{x}(\ell)\right\} + \\ &+ E\left\{\mathbf{v}^H(\ell)\mathbf{E}^H_{\hat{\mathbf{H}}\hat{\mathbf{h}}_{n_0}}\mathbf{E}_{\hat{\mathbf{H}}\hat{\mathbf{h}}_{n_0}}\mathbf{v}(\ell)\right\}, \quad (9b) \end{aligned}$$

respectively. Considering the possible increase in the noise power $\Delta = E\left\{\mathbf{v}^H(\ell)\mathbf{E}^H_{\hat{\mathbf{H}}\hat{\mathbf{h}}_{n_0}}\mathbf{E}_{\hat{\mathbf{H}}\hat{\mathbf{h}}_{n_0}}\mathbf{v}(\ell)\right\} - E\left\{\mathbf{v}^H(\ell)\mathbf{v}(\ell)\right\}$ and under a plausible assumption of a sufficiently large SNR, i.e., $E\left\{x^*_{n_0}(\ell)\mathbf{h}^H_{n_0}\mathbf{h}_{n_0}x_{n_0}(\ell)\right\} > \Delta$, we deduce that the power of $\mathbf{z}(\ell)$ is higher than that of $\mathbf{y}(\ell;\mathbf{h}_{n_0})$.

Of course, $\mathbf{H}$ and $\mathbf{h}_{n_0}$ are unavailable. However, assuming that an estimator $\hat{\mathbf{H}}$ is available, we can utilize the above observation to formulate an optimization problem seeking an oblique projection $\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}$ that minimizes the power of the projected measurements:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left\{E\left\{\mathbf{z}^H(\ell)\mathbf{E}^H_{\hat{\mathbf{H}}\boldsymbol{\theta}}\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}\mathbf{z}(\ell)\right\}\right\}, \quad (10a)$$

$$\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}} = \hat{\mathbf{H}}\left(\hat{\mathbf{H}}^H\mathbf{P}^{\perp}_{\boldsymbol{\theta}}\hat{\mathbf{H}}\right)^{-1}\hat{\mathbf{H}}^H\mathbf{P}^{\perp}_{\boldsymbol{\theta}}, \quad (10b)$$

$$\mathbf{P}^{\perp}_{\boldsymbol{\theta}} = \mathbf{I} - \boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H. \quad (10c)$$

As previously shown, applying $\mathbf{E}_{\hat{\mathbf{H}}\hat{\boldsymbol{\theta}}}$ to the array measurements $\mathbf{z}(\ell)$ results in a null steering toward $\mathbf{h}_{n_0}$. However, it is straightforward to validate that, in general, $\hat{\boldsymbol{\theta}}$ can be any vector from a subspace $\langle \mathbf{h}_{n_0}\mathbf{A} \rangle$, where $\langle \mathbf{A} \rangle$ is the subspace orthogonal to both $\left\langle \hat{\mathbf{H}} \right\rangle$ and $\langle \mathbf{h}_{n_0} \rangle$. In order to identify $\mathbf{h}_{n_0}$ by the application of (10a), $\langle \mathbf{A} \rangle$ should be an empty subspace. However, in most practical cases, the rank of $\hat{\mathbf{H}}$ is smaller than $M-1$, i.e. the number of active speakers in the room is smaller than the number of microphones minus one, hence $\langle \mathbf{A} \rangle$ is usually nonempty. Accordingly, in the case where the rank of $\hat{\mathbf{H}}$ is lower than $M-1$, it should be artificially increased prior to the application of (10a). Hence, when required, we substitute $\hat{\mathbf{H}}$ in (10b) by $\hat{\mathbf{H}}^c$. The rank of $\hat{\mathbf{H}}^c$ is set to $M-1$ by concatenating the previously estimated RTFs in $\hat{\mathbf{H}}$ with

randomly generated independent vectors $\{\mathbf{r}_i\}_{i=1}^{M-n_0}$:

$$\hat{\mathbf{H}}^c = \left[\hat{\mathbf{h}}_1, \cdots, \hat{\mathbf{h}}_{n_0-1}, \mathbf{r}_1, \cdots, \mathbf{r}_{M-n_0}\right]. \tag{11}$$

Let us substitute $\hat{\mathbf{H}}^c$ in (10b) and calculate the range and the null space of $\mathbf{E}_{\hat{\mathbf{H}}^c\boldsymbol{\theta}}$, using (5b) and (5a):

$$\mathbf{E}_{\mathbf{H}^c\mathbf{h}}\mathbf{H}^c = \mathbf{H}^c\left(\mathbf{H}^{cH}\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}^c\right)^{-1}\mathbf{H}^{cH}\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}^c = \mathbf{H}^c, \tag{12a}$$

$$\mathbf{E}_{\mathbf{H}^c\mathbf{h}}\mathbf{h} = \mathbf{H}^c\left(\mathbf{H}^{cH}\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{H}^c\right)^{-1}\mathbf{H}^{cH}\mathbf{P}_{\mathbf{h}}^{\perp}\mathbf{h} = \mathbf{0}. \tag{12b}$$

Thus, by replacing $\hat{\mathbf{H}}$ in (10b) with $\hat{\mathbf{H}}^c$ we manipulate only the range of the resulting oblique projection, while the null space left unmodified. Explicitly, an application of $\mathbf{E}_{\hat{\mathbf{H}}^c\boldsymbol{\theta}}$ to the received signal results in a distortionless response towards all columns of $\hat{\mathbf{H}}^c$. However, since $[\mathbf{r}_1, \cdots, \mathbf{r}_{M-n_0}]$ columns are randomly generated vectors, we expect no signals in the environment to impinge on the microphones with an array manifold equal to $[\mathbf{r}_1, \cdots, \mathbf{r}_{M-n_0}]$. It should be stressed that in case one of the vectors $[\mathbf{r}_1, \cdots, \mathbf{r}_{M-n_0}]$ is randomized such that it is parallel to $\mathbf{h}_{n_0}$ the proposed algorithm will fail. However, in practice, the probability to randomly guess a vector that is parallel to a specific RTF is very low, thus we neglect such an event. To summarize, by replacing $\hat{\mathbf{H}}$ in (10b) with $\hat{\mathbf{H}}^c$, the range of the projection is modified to include RTFs, which do not contribute any energy to the received signal. Accordantly, all the above derivations and conclusions hold.

Solving (10a) with $\hat{\mathbf{H}}^c$ instead of $\hat{\mathbf{H}}$ results in $\hat{\boldsymbol{\theta}}$ being parallel to $\mathbf{h}_{n_0}$, i.e., $\hat{\boldsymbol{\theta}} \approx \delta\mathbf{h}_{n_0}$, where δ is an arbitrary gain. By definition, the first entry of the RTF $\mathbf{h}_{n_0}$ is equal to 1, and hence, the estimator $\hat{\boldsymbol{\theta}}$ can be normalized to obtain an estimate of the RTF:

$$\hat{\mathbf{h}}_{n_0} = \frac{\hat{\boldsymbol{\theta}}}{\hat{\theta}_1}, \tag{13}$$

where $\hat{\theta}_1$ is the first element of $\hat{\boldsymbol{\theta}}$.

Finding a closed-form expression for $\hat{\boldsymbol{\theta}}$ that solves (10a) is a cumbersome task. However, we are able to derive an analytical expression for the first-order derivative of the target function $J = E\left\{\mathbf{y}^H(\ell;\boldsymbol{\theta})\mathbf{y}(\ell;\boldsymbol{\theta})\right\}, \mathbf{y}(\ell;\boldsymbol{\theta}) = \mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}\mathbf{z}(\ell)$ with respect to (w.r.t.) the mth entry of $\boldsymbol{\theta}$, θ_m. Hence, we can optimize J by applying a gradient descent search method, which results in the following iterative rule:

$$\hat{\theta}_m^i = \hat{\theta}_m^{i-1} + \mu\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_m}, \quad m = 1, \cdots, M \tag{14}$$

with μ being the step-size and i the iteration index. The gradient term is obtained by calculating the derivative of $J(\boldsymbol{\theta})$ w.r.t. to θ_m and is given by (see Appendix A):

$$\frac{\partial J(\boldsymbol{\theta})}{\partial\theta_m} = \mathrm{Tr}\left(\boldsymbol{\Psi}_m\mathbf{R} + \tilde{\boldsymbol{\Psi}}_m\mathbf{R}^H\right), \tag{15}$$

where $\mathbf{R} = E\left\{\mathbf{z}(\ell)\mathbf{y}^H(\ell;\boldsymbol{\theta})\right\}$ is the cross correlation matrix between the measurements and the projected signal $\mathbf{y}(\ell;\boldsymbol{\theta})$.

Algorithm 1: BOP-SRI with active speakers counting

Initialization:
A. Utilize frames $0 < \ell \leq \ell_1$ to compute $\hat{\boldsymbol{\Phi}}_{\mathrm{vv}}$ using (17).
B. Set $\Delta\mathrm{EV}_{\mathrm{Th}}$. (defined in Section IV-B)
C. Set $\mathrm{T_h}$ (defined in Section IV-A).
For each frame $\mathbf{z}(\ell)$**:**
1. Count the number of active sources (Section IV-B):
i. $\mathbf{z}_w(\ell) = \hat{\boldsymbol{\Phi}}_{\mathrm{vv,L}}^{-1}\mathbf{z}(\ell)$.
ii. $\boldsymbol{\Phi}_{z_w z_w}(\ell) = \gamma\boldsymbol{\Phi}_{z_w z_w}(\ell-1) + \mathbf{z}_w(\ell)\mathbf{z}_w^H(\ell), \gamma < 1$.
iii. Compute eigenvalue decomposition of $\boldsymbol{\Phi}_{z_w z_w}(\ell)$.
iv. $\mathrm{AS}(\ell) =$ number of eigenvalues larger than $EV_{TH}(\ell)$.

2. if $\mathrm{AS}(\ell) = \mathrm{AS}(\ell-1)$.
i. Go back to 1.

3. if $AS(\ell) < AS(\ell-1)$.
i. For $i = 1, \ldots, AS(\ell-1)$, search for $\hat{x}_i(\ell,k)$ with the least energy level and update $\hat{\mathbf{H}}$ accordingly.
ii. Go back to 1.

4. if $AS(\ell) > AS(\ell-1)$.
i. For $m = 1, \ldots, M$, set $\hat{\theta}_m^0$ to a small random number.
ii. Apply (17) for $m = 1, \ldots, M$, till convergence.
iii. Monitor local minimum (Section IV-A).
vi. Apply (13).
v. Output $\hat{\mathbf{h}}_{n_0}$ and go back to 1.

The rest of the terms are defined as

$$\boldsymbol{\Psi}_m = \hat{\mathbf{H}}\boldsymbol{\Gamma}\boldsymbol{\Omega}_m\left(\hat{\mathbf{H}}\boldsymbol{\Gamma}\mathbf{P}_{\boldsymbol{\theta}}^{\perp} - \mathbf{I}\right), \tag{16a}$$

$$\tilde{\boldsymbol{\Psi}}_m = \left(\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\boldsymbol{\Gamma}^H\hat{\mathbf{H}}^H - \mathbf{I}\right)\boldsymbol{\Omega}_m\boldsymbol{\Gamma}^H\hat{\mathbf{H}}^H, \tag{16b}$$

$$\boldsymbol{\Gamma} = \left(\hat{\mathbf{H}}^H\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\hat{\mathbf{H}}\right)^{-1}\hat{\mathbf{H}}^H, \tag{16c}$$

$$\boldsymbol{\Omega}_m = \left(\boldsymbol{\theta}^{\dagger}\right)^H\mathbf{i}_m - \theta_m\left(\boldsymbol{\theta}^{\dagger}\right)^H\boldsymbol{\theta}^{\dagger}, \tag{16d}$$

where $(\cdot)^{\dagger}$ is the pseudo inverse operator and $\mathbf{i}_m$ is a vector with its mth element equal to 1 and the rest are zeros.

IV. Practical Considerations

The proposed, BOP-SRI method is presented in Algorithm 1. However, several practical aspects related to the algorithm implementation should be considered.

A. Initialization

Since in the general case the cost function $J(\boldsymbol{\theta})$ is not unimodal, the iterative method (14) could become trapped in a local minimum, depending on the initial conditions. It is therefore important to initialize the search algorithm in close proximity to the global minimum. We are, however, not familiar with an easy method for computing a good starting point for the iterative optimization problem at hand. Accordingly, in the scope of this work we followed the following procedure. We initialized $\hat{\theta}_m^0, m = 1, \ldots, M$ with randomly generated complex-valued numbers. During the convergence, we monitored the value of the cost function $J(\boldsymbol{\theta})$. When the difference between the value of the cost function at the final iteration and its initial value was within a predefined threshold $\mathrm{T_h}$, we discarded this optimization cycle and re-initialized $\hat{\theta}_m^0$. This initialization procedure was proven efficient in terms of the resulting RTF estimation quality, as presented in Section V.

However, in terms of the resulting computational load, the proposed procedure may be suboptimal. A better initialization method is, however, beyond the scope of the current contribution.

B. Activity indicator function

The proposed method assumes that the activity indicator function of the sources of interest $\mathcal{I}(\ell)$ is available to the algorithm. The BOP-SRI procedure utilizes $\mathcal{I}_{n_0}(\ell)$ to address the challenge induced by a *birth* of a speaker. In a practical scenario, $\mathcal{I}_{n_0}(\ell)$ should be deduced from the measurements $\mathbf{z}(l,k)$. In addition, an RTF *death* mechanism is also required. Refer to [23] for an equivalent discussion in dynamic scenarios. Source counting methods [37] may be useful for detecting the number of active sources in a specific time period. Since a simultaneous *birth* and *death* of two independent speakers seldom occurs, the BOP-SRI process is triggered when an increase in the number of active sources occurs. An RTF *death* mechanism may be triggered when a decrease in the number of active speakers occurs. For example, the ith RTF may be considered obsolete if the power of $\hat{x}_i(\ell,k)$ is below a threshold for a predetermined period of time. That being said, in practical situations, where speakers may arbitrary start and stop speaking, an RTF association mechanism [38] is likely to be required. However, such a mechanism is beyond the scope of the current work.

In the scope of this work, we employed an active source counting method based on the microphone signals PSD matrix generalized eigenvalue decomposition (EVD) [16]. Namely, the PSD matrix of the stationary noise $\mathbf{\Phi}_{\text{vv}}$ is estimated during speech absent periods, by a sample covariance estimator:

$$\widehat{\mathbf{\Phi}}_{\text{vv}} = \frac{1}{\ell_1}\sum_{\ell=1}^{\ell_1}\mathbf{z}(\ell)\mathbf{z}^H(\ell). \tag{17}$$

Then, the microphone signals are whitened using $\mathbf{z}_w(\ell) = \hat{\mathbf{\Phi}}_{\text{vv,L}}^{-1}\mathbf{z}(\ell)$, where $\hat{\mathbf{\Phi}}_{\text{vv,L}}$ is the lower triangular matrix obtained by the Cholesky decomposition of the stationary noise PSD matrix estimate. Let $\mathbf{\Phi}_{z_w z_w}$ be a PSD matrix of the whitened measurements, using EVD we have $\mathbf{\Phi}_{z_w z_w} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1}$, where $\mathbf{E}$ is a square matrix with columns corresponding to the eigenvectors of $\mathbf{\Phi}_{z_w z_w}$ and $\mathbf{\Lambda}$ is a diagonal matrix, which diagonal elements are the corresponding eigenvalues of $\mathbf{\Phi}_{z_w z_w}$. The signal $\mathbf{z}_w(\ell)$ consists of components contributed by the active speakers in the environment and a white noise. Hence, assuming the number of microphones is larger than the number of speakers, the larger eigenvalues can be attributed to the coherent signals (speech) while the lower to the spatially white signals. The number of active speech sources is then inferred by counting the number of elements in $\mathbf{\Lambda}$ that are above certain threshold. In a practical scenario, considering the whitening and modelling errors, the threshold is set to $\text{EV}_{\text{Th}}(\ell) = \lambda_{min}(\ell) + \Delta\text{EV}_{\text{Th}}$. Where, $\lambda_{min}(\ell)$ is the lowest eigenvalue, and $\Delta\text{EV}_{\text{Th}}$ is a predefined constant. The applicability of aforementioned active source counting method is demonstrated in Section V-D.

V. Experimental Study

We turn now to the evaluation of the performance of the proposed BOP-SRI algorithm. The evaluation was split into two experiments. In the first experiment, we aimed at evaluating the BOP-SRI performance for various SNR and signal to interference ratio (SIR) scenarios. The second experiment was dedicated to exploring the speech separation performance of an LCMV beamformer computed based on the RTFs estimated by the BOP-SRI algorithm.

A. Setup and definitions

The proposed BOP-SRI algorithm was tested using a multichannel impulse response database (MIRDB)[1] measured in the speech and acoustic laboratory of the Faculty of Engineering at Bar-Ilan University [39]. The laboratory is a $6 \times 6 \times 2.4$ m room with variable reverberation times. The database consists of impulse responses relating eight microphones arranged to form a linear array and the loudspeaker placed at various angles from -90° to 90° at distances of 1 and 2 m from the array. The processing was executed in the frequency domain, the STFT analysis window length was set to 4096, with 75% overlap between successive frames, while the sampling frequency of the system was set to 16 KHz. The performance of the BOP-SRI algorithms was manifested by the blocking ability factor (BAF)

$$\text{BAF}_n \triangleq \frac{1}{M-1}\sum_{m=2}^{M}\frac{\sigma_{m,n}^2}{\sigma_{m,\text{v}}^2}\frac{E\left\{\left[v_m(t)-\widehat{h}_{m,n}(t)*v_1(t)\right]^2\right\}}{E\left\{\left[x_{m,n}(t)-\widehat{h}_{m,n}(t)*x_{1,n}(t)\right]^2\right\}},$$

where $x_{m,n}(t)$ is the speech generated by $s_n(t)$ and measured by the mth microphone, $v_m(t)$ is the noise at the mth microphone, $\sigma_{m,n}^2$ is the power of $x_{m,n}(t)$, $\sigma_{m,\text{v}}^2$ is the power of $v_m(t)$, $\widehat{h}_{m,n}(t)$ is the estimated RTF relating the first and the mth microphone as a response to $s_n(t)$, and $E\{[\cdot]^2\}$ is the power of $[\cdot]$. The blocking ability factor BAF_n measures the ratio between the ability to block the nth speech source and its inherent ability to block a random noise. BAF has a major effect on the amount of distortion introduced by the transfer function GSC due to desired speech leakage [10].

B. BOP-SRI performance vs. SNR and SIR

We turn now to the evaluation of the BOP-SRI algorithm's performance in various SNR and SIR conditions. We utilized the MIRDB to setup a uniform linear array comprising $M = 3$ microphones with 3 cm inter-spacing, capturing a mixture of three acoustic sources positioned at a distance of 1 m from the array, while T_{60} was set to 160 mSec. Specifically, two speech sources $s_1(t)$ and $s_2(t)$ impinged the array with angle of arrivals (AOAs) equal to 0° and 30°, respectively, and a stationary fan noise $v(t)$ impinged the array with AOA equal to 300°. The powers of the sources were defined as σ_1^2, σ_2^2 and σ_{v}^2, respectively. The activity pattern of the sources was set such that $s_1(t)$ was activated 10 sec after the start of the

[1] http://www.eng.biu.ac.il/gannot/downloads/

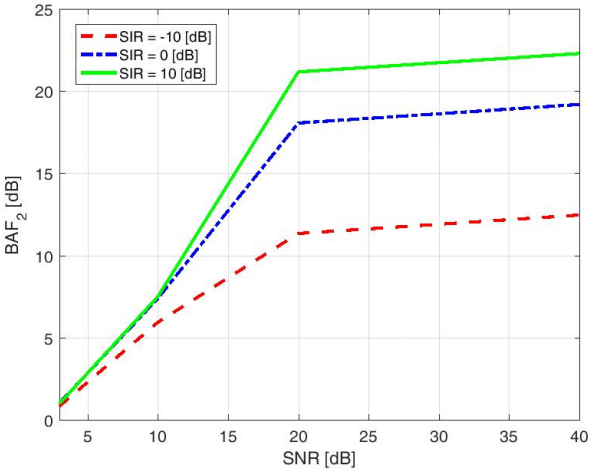

Fig. 4: Blocking ability factor of $\hat{\mathbf{h}}_2(k)$. $\hat{\mathbf{h}}_2(k)$ was estimated using BOP-SRI and time instances where $\mathcal{I}_1(\ell) = \mathcal{I}_2(\ell) = 1$. To implement the BOP-SRI, we utilized $\hat{\mathbf{h}}_1(k)$, that was estimated by CW method, during time instances where $\mathcal{I}_1(\ell) = 1, \mathcal{I}_2(\ell) = 0$.

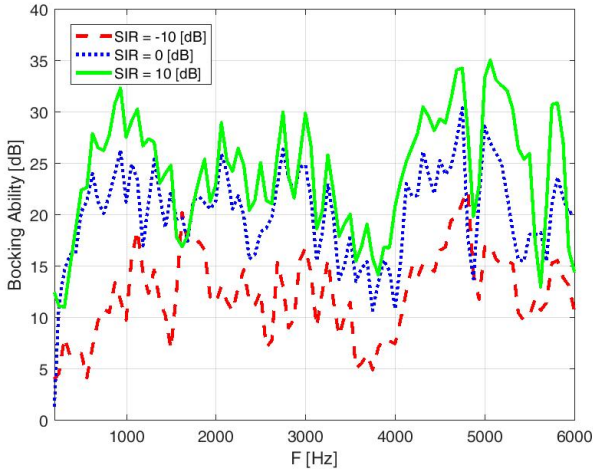

Fig. 5: Blocking ability frequency response of $\hat{\mathbf{h}}_2(k)$. $\hat{\mathbf{h}}_2(k)$ was estimated using BOP-SRI and time instances where $\mathcal{I}_1(\ell) = \mathcal{I}_2(\ell) = 1$. To implement the BOP-SRI, we utilized $\hat{\mathbf{h}}_1(k)$, that was estimated by CW method, during time instances where $\mathcal{I}_1(\ell) = 1, \mathcal{I}_2(\ell) = 0$.

measurement, the time difference between $s_1(t)$ and $s_2(t)$ activation was 10 sec, and following $s_2(t)$ activation both sources remained active for an additional 10 sec.

Our main goal in this experiment was to estimate the RTF of the second source $\mathbf{h}_2(k)$ in various SNR $= \sigma_2^2/\sigma_{\mathrm{v}}^2$ and SIR $= \sigma_2^2/\sigma_1^2$ conditions, while utilizing only the time frames where both speech sources are active $\ell_2 < \ell$. In order to accomplish this task, we computed the PSD matrix of the noise $\hat{\mathbf{\Phi}}_{\mathrm{vv}}(k)$ by utilizing the noise only time frames $0 < \ell \leq \ell_1$ followed by the application of the CW RTF estimator [16] to the signals received during noise plus first speaker active frames, namely, $\ell_1 < \ell \leq \ell_2$. The CW resulted in an estimate of the first speaker RTF $\hat{\mathbf{h}}_1(k)$. We then artificially increased the rank of the matrix $\hat{\mathbf{H}}(k)$ by appending an arbitrary vector $\mathbf{r}$, namely, $\hat{\mathbf{H}}^c(k) = \left[\hat{\mathbf{h}}_1(k)\ \mathbf{r}\right]$. The BOP-SRI algorithm was applied by implementing (14) and (15) with $\hat{\mathbf{H}}^c(k)$ (for each frequency bin k) until convergence. The threshold $\mathrm{T_h}$ was set to 10 dB and the step size μ was set to 0.1.

The resulting BAF of $\hat{\mathbf{h}}_2(k)$ for various SNR and SIR values is depicted in Fig. 4. As can be readily seen, both the SNR and SIR influence the BOP-SRI estimation accuracy. It is also seen that increasing the SNR above 20 dB for each SIR results in a minor improvement in the estimation accuracy, whereas reducing the SNR below 20 dB greatly affects the performance. Considering the SIR influence, it seems that reducing the SIR from 0 dB to -10 dB has a greater effect on the estimation performance than an SIR reduction from 10 dB to 0 dB.

In Fig. 5, the blocking ability of $\hat{\mathbf{h}}_2(k)$ is presented as a function of the frequency for various SIR values while the SNR is set to 20 dB. As expected, the resulting blocking ability is very nonuniform across the frequency range. For example, the blocking is relatively poor at the low frequencies, for all the considered SIRs. The nonuniform blocking can be attributed to the spectral characteristic of the speech signals.

In Fig. 6, we exemplify the iterative minimization processes (14) by depicting the evolvement in the components of the cost function

$$J_1 = E\left\{\mathbf{x}_1^H(\ell)\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}^H\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}\mathbf{x}_1(\ell)\right\},$$
$$J_2 = E\left\{\mathbf{x}_2^H(\ell)\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}^H\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}\mathbf{x}_2(\ell)\right\},$$
$$J_{\mathrm{v}} = E\left\{\mathbf{v}^H(\ell)\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}^H\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}\mathbf{v}(\ell)\right\},$$

for a single frequency, $f = 1500$ Hz. As can be seen, the optimization is manifested through a minimization of J_2 until a convergence is reached after ≈ 2300 iterations. It is also seen that, as expected, J_1 remains constant throughout the minimization processes, since $\hat{\mathbf{h}}_1(k)$ is in the range $\mathbf{E}_{\hat{\mathbf{H}}^c\boldsymbol{\theta}}^H$. Fig. 6 also demonstrates the noise enhancement phenomenon by $\mathbf{E}_{\hat{\mathbf{H}}^c\boldsymbol{\theta}}^H$; this is manifested through the J_v increase during the iterative process. This stresses again the SNR's importance to the proposed BOP-SIR algorithm. In this specific example, BOP-SIR is aimed at minimizing the cost function $J = J_1 + J_2 + J_v$. Accordingly, any decrease in J_2 may be masked by an increase in J_v; to prevent this masking effect, the level of the noise $\mathbf{v}$ should be sufficiently low as compared with the level of the signal $\mathbf{x}_2$.

C. Speech extraction

In this section, we demonstrate the effectiveness of the proposed BOP-SRI algorithm in a beamforming application. Our objective was to estimate the RTFs in a multi-speaker scenario and then apply an LCMV beamformer for extracting the individual speakers from the measured mixture.

We utilized the MIRDB to implement a uniform linear array comprising $M = 8$ microphones, capturing a mixture of four acoustic sources, while T_{60} was selected to be 160 mSec. Specifically, three equipower speech sources $s_1(t), s_2(t), s_3(t)$ were generated; $s_1(t)$ and $s_2(t)$ were positioned at a distance of 1 m from the array, with AOAs equal to 0° and 45°,

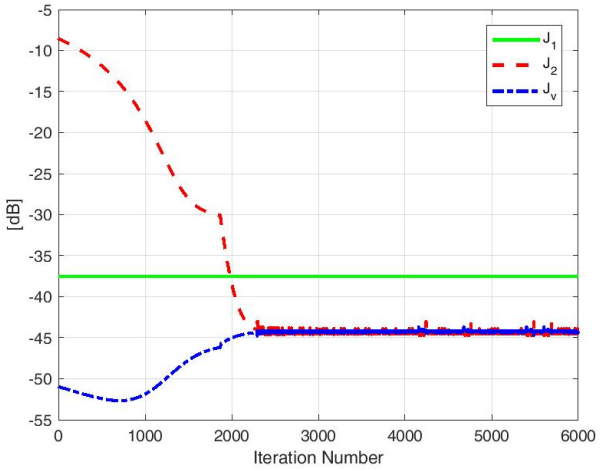

Fig. 6: Responses of the oblique projection $\mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}$ toward $\mathbf{x}_1(\ell), \mathbf{x}_2(\ell)$ and $\mathbf{v}(\ell)$.

respectively. $s_3(t)$ was positioned at a distance of 2 m from the array, with AOAs equal to $315°$. A stationary noise source $v(t)$ was positioned at a distance of 2 m with AOA equal to $90°$, and the SNR was set to 20 dB and the noise was active throughout the experiment.

In the first part of the experiment, all the speech sources were inactive, and accordingly, signals measured during this part of the experiment were utilized for estimating the noise PSD matrix. In the second part, the speech sources were activated in a non-overlapping manner, each for a period of 10 sec. The signals measured during this part of the experiment were utilized for estimating the RTFs of $s_1(t), s_2(t), s_3(t)$. The RTFs were estimated by applying the well established CW RTF estimator [16], which is a valid RTF estimator in a noisy environment with a single active speech source. These RTFs are referred to in the following as $\hat{\mathbf{h}}_1^{\text{EVD}}, \hat{\mathbf{h}}_2^{\text{EVD}}, \hat{\mathbf{h}}_3^{\text{EVD}}$, respectively. In the third part, both $s_1(t)$ and $s_2(t)$ were active during the first 10 sec, while during the following 10 sec all the speech source were concurrently active. The signals measured during the first 10 sec and $\hat{\mathbf{h}}_1^{\text{EVD}}$ were utilized for estimating the RTF of $s_2(t)$ by applying the proposed BOP-SRI algorithm. This RTF is referred to in the following as $\hat{\mathbf{h}}_2^{\text{BOP}}$. The signals measured during the next 10 sec, as well as $\hat{\mathbf{h}}_1^{\text{EVD}}$ and $\hat{\mathbf{h}}_2^{\text{BOP}}$, were utilized for estimating the RTF of $s_3(t)$ by applying the proposed BOP-SRI algorithm. This RTF is referred to in the following as $\hat{\mathbf{h}}_3^{\text{BOP}}$.

The above mentioned estimators facilitated the implementation of an LCMV beamformer aimed at extracting the desired speech source from the measurements. The beamformers are referred to in the following as $\mathbf{w}_n^{\text{est}}$, where est $\in \{\text{EVD}, \text{BOP}, \text{DOA}\}$, and $n \in \{1, 2, 3\}$. The constraints set of an LCMV beamformer marked with a superscript EVD was formulated by utilizing $\left[\hat{\mathbf{h}}_1^{\text{EVD}}, \hat{\mathbf{h}}_2^{\text{EVD}}, \hat{\mathbf{h}}_3^{\text{EVD}}\right]$, while a superscript BOP indicates that the constraints of the beamformer were formulated by utilizing $\left[\hat{\mathbf{h}}_1^{\text{EVD}}, \hat{\mathbf{h}}_2^{\text{BOP}}, \hat{\mathbf{h}}_3^{\text{BOP}}\right]$. A superscript DOA means that the constraints of the beamformer were formulated using directional array manifolds steered towards the known DOA of the sources. A subscript n in $\mathbf{w}_n^{\text{est}}$ indicates a beamformer with distortionless response to $\mathbf{x}_n$ and zero response to the other two speech sources in the room, where $\mathbf{x}_n$ is the image of s_n as measured by the microphones.

TABLE I: Linearly constrained minimum variance beamformer gains toward the sources of interest

	$\mathbf{x}_1$ gain [dB]	$\mathbf{x}_2$ gain [dB]	$\mathbf{x}_3$ gain [dB]
$\mathbf{w}_1^{\text{EVD}}$	0.22	-22.22	-17.95
$\mathbf{w}_1^{\text{BOP}}$	0.11	-14.78	-11.05
$\mathbf{w}_1^{\text{DOA}}$	-0.8	-2.77	-1.39
$\mathbf{w}_2^{\text{EVD}}$	-22.48	0.15	-17.78
$\mathbf{w}_2^{\text{BOP}}$	-24.36	-1.03	-10.65
$\mathbf{w}_2^{\text{DOA}}$	-5.5	-2.13	-0.1
$\mathbf{w}_3^{\text{EVD}}$	-23.77	-24.04	0.25
$\mathbf{w}_3^{\text{BOP}}$	-26.28	-16.94	0
$\mathbf{w}_3^{\text{DOA}}$	-4.45	-2.64	-1.47

It should be stressed again that, although in the following we compare the performance of the proposed BOP-SRI algorithm to that of the well-established CW algorithm, the algorithms address two different challenges. The BOP-SRI addresses RTF estimation in a noisy and multi-speaker scenario, whereas the CW method addresses RTF estimation in a noisy single-speaker scenario. In a multi-speaker scenario, the CW method results in the identification of orthogonal vectors that span the sub-space of all the active speakers in the environment. These orthogonal vectors are different from the individual RTFs and, to the best of our knowledge, there is no established method available for inferring the individual RTFs from the orthogonal vectors. Since the CW addresses an easier task, it is used in the following as a bound for the BOP-SRI performance. In order to demonstrate the advantage of using BOP-SRI in multi-speaker scenario we also present a speech separation performance results by applying a directional beamformer.

The resulting gains of the LCMV beamformers toward the sources of interest are summarized in Table I. The presented scalar gains are the averaged results over all frequency bands. It can be easily verified that the performance of $\mathbf{w}_n^{\text{BOP}}, n = 1, 2, 3$ beamformers is inferior to that of the beamformers $\mathbf{w}_n^{\text{EVD}}, n = 1, 2, 3$. However, the performance of $\mathbf{w}_n^{\text{BOP}}, n = 1, 2, 3$ beamformers is still reasonably high, significantly higher compared to the directional beamformer calculated using known DOAs. It should be noted that the gain of $\mathbf{w}_1^{\text{BOP}}$ towards $\mathbf{x}_1$ differs from the respective gain of $\mathbf{w}_1^{\text{EVD}}$. While $\mathbf{w}_1^{\text{EVD}}$ beamformer is calculated using $\hat{\mathbf{h}}_1^{\text{EVD}}, \hat{\mathbf{h}}_2^{\text{EVD}}, \hat{\mathbf{h}}_3^{\text{EVD}}$, the $\mathbf{w}_1^{\text{BOP}}$ beamformer is calculated using $\hat{\mathbf{h}}_1^{\text{EVD}}, \hat{\mathbf{h}}_2^{\text{BOP}}, \hat{\mathbf{h}}_3^{\text{BOP}}$. This leads to the difference in beamformers' gains. At this point we would like to stress again that the EVD beamformers are used in this comparison as an unrealistic bound, as estimation of these beamformers requires an oracle scenario where the sources are active in a non-overlapping manner.

In Fig. 7, Fig. 8, and Fig. 9, we present the frequency responses of the considered beamformers $\mathbf{w}_n^{\text{est}}$, est $\in \{\text{EVD}, \text{BOP}\}, n \in \{1, 2, 3\}$, toward the speech sources of interest, $\mathbf{x}_n, n = 1, 2, 3$. The figures suggest that BOP-SRI resulted in a better estimation of the RTF of $\mathbf{x}_2$ than the RTF of $\mathbf{x}_3$. This can be attributed to the fact that s_3 is farther

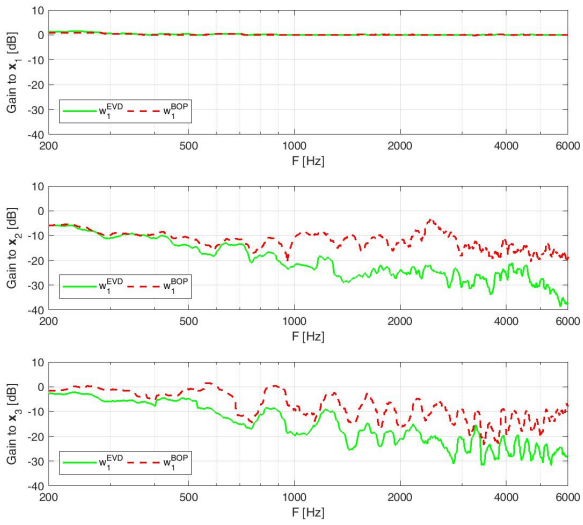

Fig. 7: Frequency responses of the $\mathbf{w}_1^{\text{est}}, \text{est} \in \{\text{EVD}, \text{BOP}\}$ beamformers toward sources of interest.

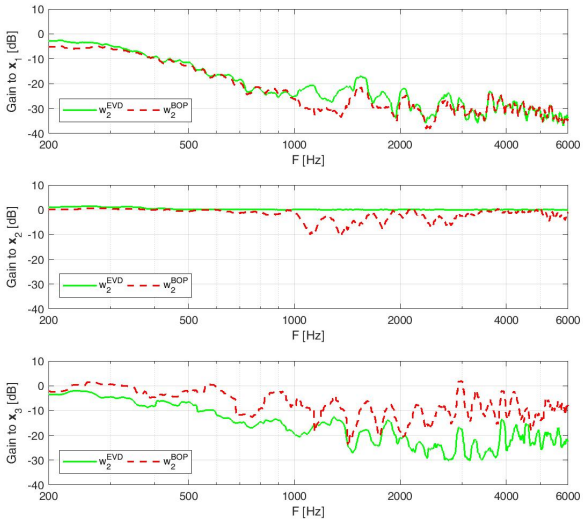

Fig. 8: Frequency responses of the $\mathbf{w}_2^{\text{est}}, \text{est} \in \{\text{EVD}, \text{BOP}\}$ beamformers toward sources of interest.

away from the microphones array than s_2, which results in a worse SNR. Additionally, the SIR was worse during the $\hat{\mathbf{h}}_3^{\text{BOP}}$ estimation than during the $\hat{\mathbf{h}}_2^{\text{BOP}}$ estimation, since during the $\hat{\mathbf{h}}_3^{\text{BOP}}$ estimation both s_1 and s_2 were active, while during the estimation of $\hat{\mathbf{h}}_2^{\text{BOP}}$ only s_1 was active.

D. Speech enhancement in a babble noise environment

In this section, we demonstrate the effectiveness of the proposed BOP-SRI algorithm in the presence of babble noise, instead of the coherent noise sources that were considered in the previous experiments. The performance of the BOP-SRI algorithm is manifested by an application of an LCMV beamformer that extracts the speaker, the RTF of which is estimated by the BOP-SRI.

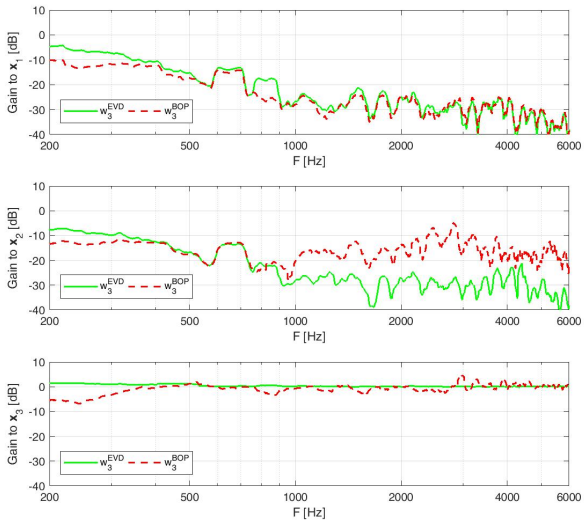

Fig. 9: Frequency responses of the $\mathbf{w}_3^{\text{est}}, \text{est} \in \{\text{EVD}, \text{BOP}\}$ beamformers toward sources of interest.

We also utilized the MIRDB to implement a uniform linear array comprising $M = 8$ microphones. In this experiment, the reverberation time in the room was higher than in the previous experiments, $T_{60} = 360$ mSec. Two equipower speech sources $s_1(t), s_2(t)$ were positioned at a distance of 1 m from the array, with AOAs equal to $0°, 45°$, respectively. Microphone signals were further corrupted by babble noise $v(t)$, which was played through four loudspeakers positioned in the room and facing the walls, with the SNR set to 8 dB. Similarly to the previous experiment, the speech sources became successively active while the noise was active throughout the experiment. However, unlike in the previous experiment, in this experiment, the activity pattern of the sources was unknown to the BOP-SRI algorithm, i.e., it was inferred by counting the number of dominant eigenvalues of the whitened measurements PSD matrix $\mathbf{\Phi}_{z_w z_w}$.

For reference, similarly to the previous experiment, the $\mathbf{w}_n^{\text{BOP}}, n = 1, 2$ performance is compared with the $\mathbf{w}_n^{\text{EVD}}, n = 1, 2$ performance. For estimating the RTFs for the $\mathbf{w}_n^{\text{EVD}}$ beamformers, each of the sources $s_1(t), s_2(t)$ was recorded separately in the presence of the babble noise.

The implementation of the active speakers counting method proposed in Section IV-B is presented first. During the period of the first 10 sec, both $s_1(t)$ and $s_2(t)$ were inactive. This period is used to estimate the PSD matrix of the stationary noise $\mathbf{\Phi}_{\text{vv}}$. Upon estimating $\mathbf{\Phi}_{\text{vv}}$, the signal received at each time frame was whitened and the eigenvalue decomposition of $\mathbf{\Phi}_{z_w z_w}$ was computed. For example, we present the eigenvalues of $\mathbf{\Phi}_{z_w z_w}$ during a time frame with a single active speaker in Fig. 10. As readily seen, a single dominant eigenvalue is demonstrated. Also, it should be noted that the other eigenvalues are different from 0 dB magnitude, this should be attributed to whitening and model errors. Equivalently, we demonstrate in Fig. 11 the eigenvalues of $\mathbf{\Phi}_{z_w z_w}$ during a time frame with two active speakers, two dominant eigenvalues are readily identified. The entire processes of active speakers

counting is depicted in Fig. 12, where the power level of each of the eigenvalues of $\mathbf{\Phi}_{z_w z_w}$ is presented as a function of the time as well as the threshold $EV_{TH}(\ell)$ with $\Delta EV_{TH} = 60$ dB. The number of active speakers in each time frame is inferred by counting the number of eigenvalue with higher magnitude than $EV_{TH}(\ell)$. Ultimately, we present the inferred speakers activity pattern in Fig. 13.

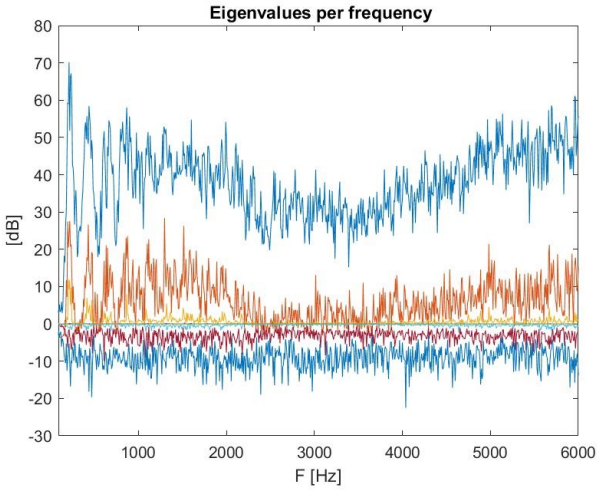

Fig. 10: Eigenvalues of single active speaker segment as a function of the frequency. A Single dominant eigenvalue is readily identified.

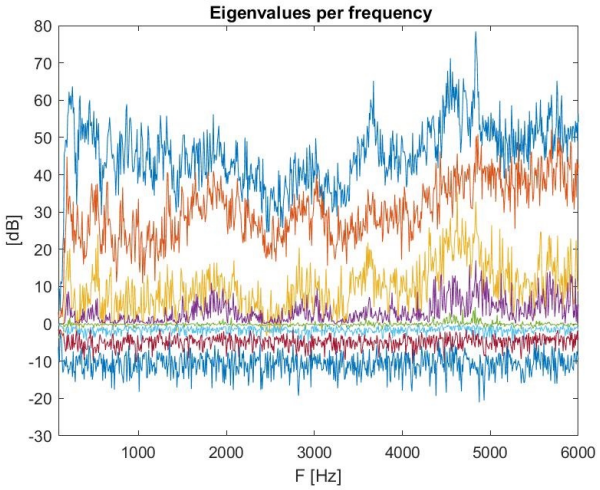

Fig. 11: Eigenvalues of two active speakers segment as a function of the frequency. Two dominant eigenvalues are readily identified.

Upon inferring the speakers activity pattern from the measurements, we carried out the speech enhancement experiment in a similar manner to the previous experiment, with a single difference. The inferred speakers activity function was used to trigger the BOP-SRI algorithm instead of the oracle function used in the previous experiment. The resulting gains of the LCMV beamformers toward the sources of interest are presented in Table II. The scalar gains are the results of frequency averaging. It can be easily concluded that, similarly

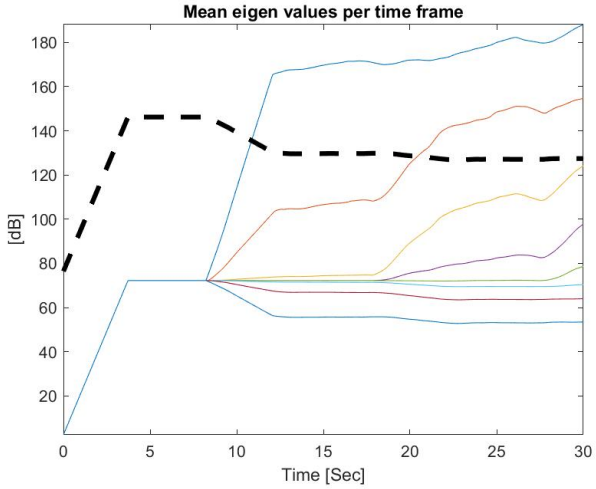

Fig. 12: Averaged eigenvalues power level as a function of time. The bold dashed line represents the $\mathrm{EV}_{\mathrm{Th}}(\ell)$. The number of active speakers is inferred by counting the number of eigenvalues with higher energy level that $\mathrm{EV}_{\mathrm{Th}}(\ell)$ during each time segment.

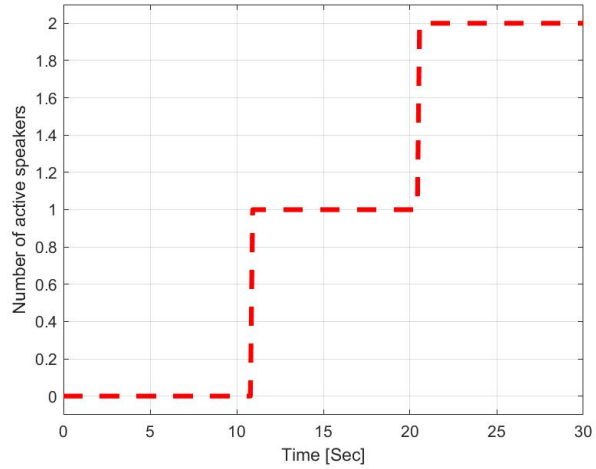

Fig. 13: Speakers activity pattern as estimated by the proposed active speakers counting method.

to the previous experiment, the $\mathbf{w}_n^{\mathrm{BOP}}, n = 1, 2$ beamformers result in a worse performance than the $\mathbf{w}_n^{\mathrm{EVD}}, n = 1, 2$ beamformers. However, the $\mathbf{w}_n^{\mathrm{BOP}}, n = 1, 2$ beamformers still perform reasonably well, although $\hat{\mathbf{h}}_2^{\mathrm{BOP}}$ was estimated in a multi-speaker scenario in the presence of babble noise.

In Fig. 14 and Fig. 15, the frequency responses of the considered beamformers $\mathbf{w}_n^{\mathrm{est}}, \mathrm{est} \in \{\mathrm{EVD}, \mathrm{BOP}\}, n \in \{1, 2\}$, toward the speech source of interest, $\mathbf{x}_n, n = 1, 2$, are presented. On top of the already formulated conclusions, these figures suggest that the proposed BOP-SRI algorithm is applicable for successive estimation of the RTFs in the presence of babble noise and without a priori knowledge of the sources' activity pattern, as was assumed in the previous experiment.

To inquire the robustness of the proposed method to reverberation we repeated this exact experiment with a single

TABLE II: Linearly constrained minimum variance beamformers' gains toward the sources of interest, $T_{60} = 360$ mSec.

	$\mathbf{x}_1$ gain [dB]	$\mathbf{x}_2$ gain [dB]
$\mathbf{w}_1^{\text{EVD}}$	0.44	-15.4
$\mathbf{w}_1^{\text{BOP}}$	0.46	-10.5
$\mathbf{w}_2^{\text{EVD}}$	-14.54	0.370
$\mathbf{w}_2^{\text{BOP}}$	-16.66	-1.70

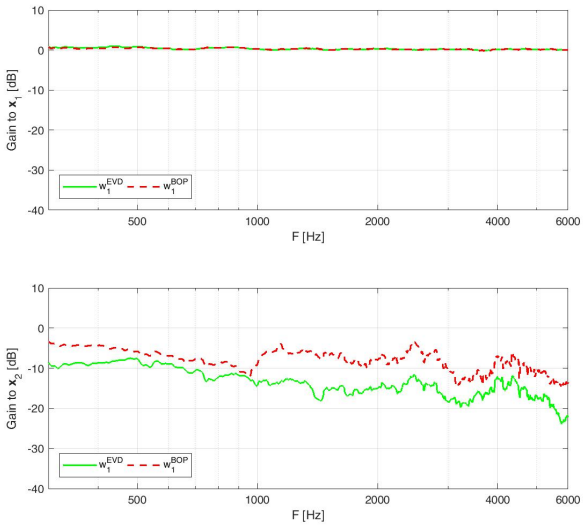

Fig. 14: Frequency responses of the $\mathbf{w}_1^{\text{est}}, \text{est} \in \{\text{EVD}, \text{BOP}\}$ beamformers toward sources of interest.

TABLE III: Linearly constrained minimum variance beamformers' gains toward the sources of interest, $T_{60} = 610$ mSec.

	$\mathbf{x}_1$ gain [dB]	$\mathbf{x}_2$ gain [dB]
$\mathbf{w}_1^{\text{EVD}}$	0.14	-14.6
$\mathbf{w}_1^{\text{BOP}}$	0.23	-11.0
$\mathbf{w}_2^{\text{EVD}}$	-15.13	0.43
$\mathbf{w}_2^{\text{BOP}}$	-17.83	1.5

difference, this time $T_{60} = 610$ mSec. The resulting gains of the LCMV beamformers toward the sources of interest are presented in Table III. The scalar gains are the results of frequency averaging. It can be easily concluded that, the resulting gains in this scenario are similar to $T_{60} = 360$ mSec scenario. This suggests that the proposed BOP-SRI is an applicable method in a reverberant scenario.

VI. Summary

In this contribution, the challenge of RTF identification in a multi-speaker scenario was considered. We introduced the SRI approach, which is based on the sole assumption that sources do not become simultaneously active. In particular, we addressed the challenge of estimating the RTF of a specific speech source while assuming that the RTFs of all the other active sources in the environment were previously estimated.

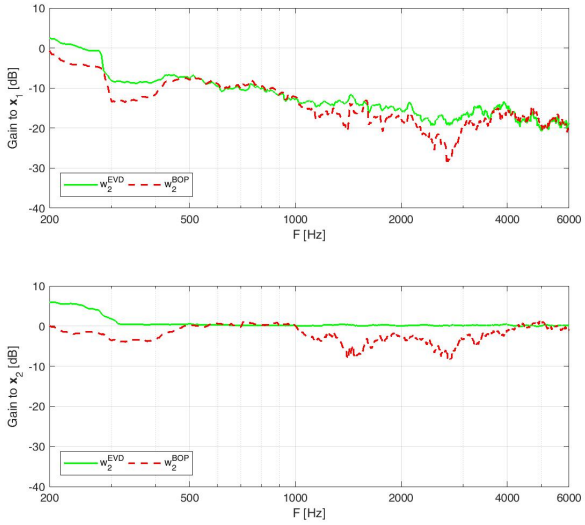

Fig. 15: Frequency responses of the $\mathbf{w}_2^{\text{est}}, \text{est} \in \{\text{EVD}, \text{BOP}\}$ beamformers toward sources of interest.

The RTF of interest was identified by applying the BOP-SRI technique. Upon the identification of a new speech source in the environment, the BOP algorithm is applied. Applying BOP results in an oblique projection matrix that, once applied to the microphone measurements, results in a null steering toward the RTF of interest. We proved that, by artificially inflating the range of the projection matrix, the RTF of interest can be inferred. We established an experimental setup based on the MIRDB, which facilitated a performance evaluation of the proposed BOP-SRI algorithm in various SNR and SIR and reverberation levels conditions. The applicability of the RTF estimated by the BOP-SRI in a multi-speaker environment was tested in a speech extraction task. We compared the performance of two sets of LCMV beamformers, where the first set was calculated by utilizing the RTF estimated in a single speaker environment by applying the CW method, while the second set was calculated by utilizing the RTFs estimated by BOP-SRI in a multi-speaker environment. Unsurprisingly, the first set of beamformers results in a better speech extraction performance than the second set. At this point we would like to stress again that the beamformers estimated using the CW method are used in this study as an unrealistic bound, as estimation of these beamformers requires an oracle scenario where the sources are active in a non-overlapping manner. However, the proposed BOP-SRI method provides the flexibility to identify an RTF in a more realistic multi-speaker environment, while still resulting in a reasonable performance in the considered experiment. The applicability of the proposed BOP-SRI algorithm was also tested in the presence of babble noise and without a priori knowledge of the sources activity pattern. It may be worth stressing that the proposed technique was only evaluated in controlled environments, both simulated and actual acoustic lab. Applying the proposed technique to more complex scenarios with arbitrary activity patterns of the speakers, may require more sophisticated speakers counting and RTF association algorithms.

APPENDIX A
GRADIENT DERIVATION

The gradient of the target function $J(\boldsymbol{\theta})$ w.r.t. the mth element of $\boldsymbol{\theta}$, $\theta_m = r_m + jc_m$ can be computed by applying the chain rule for a complex valued function, resulting in

$$\begin{aligned}\frac{\partial}{\partial \theta_m} E\left\{\mathbf{y}^H(\ell;\boldsymbol{\theta})\mathbf{y}(\ell;\boldsymbol{\theta})\right\} &= E\left\{\mathbf{y}^H(\ell;\boldsymbol{\theta})\frac{\partial \mathbf{y}(\ell;\boldsymbol{\theta})}{\partial r_m}\right\} + \\ &+ E\left\{\left(\frac{\partial \mathbf{y}(\ell;\boldsymbol{\theta})}{\partial r_m}\right)^H \mathbf{y}(\ell;\boldsymbol{\theta})\right\} + jE\left\{\mathbf{y}^H(\ell;\boldsymbol{\theta})\frac{\partial \mathbf{y}(\ell;\boldsymbol{\theta})}{\partial c_m}\right\} \\ &+ jE\left\{\left(\frac{\partial \mathbf{y}(\ell;\boldsymbol{\theta})}{\partial c_m}\right)^H \mathbf{y}(\ell;\boldsymbol{\theta})\right\}. \end{aligned} \tag{18}$$

Let us write the explicit form of the projected vector $\mathbf{y}(\ell;\boldsymbol{\theta})$ by utilizing (10b) and (10c):

$$\begin{aligned}\mathbf{y}(\ell;\boldsymbol{\theta}) &= \mathbf{E}_{\hat{\mathbf{H}}\boldsymbol{\theta}}\mathbf{z}(\ell) = \\ &= \hat{\mathbf{H}}\left(\hat{\mathbf{H}}^H\left(\mathbf{I} - \boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H\right)\hat{\mathbf{H}}\right)^{-1}\hat{\mathbf{H}}^H \times \\ &\quad \left(\mathbf{I} - \boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H\right)\mathbf{z}(\ell). \end{aligned} \tag{19}$$

The derivative of $\mathbf{y}(\ell)$ w.r.t. to θ_m is computed straightforwardly; after collecting like terms, it results in the expression

$$\frac{\partial}{\partial \theta_m}\mathbf{y}(\ell;\boldsymbol{\theta}) = \hat{\mathbf{H}}\boldsymbol{\Gamma}\frac{\partial}{\partial \theta_m}\left[\boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H\right]\left(\hat{\mathbf{H}}\boldsymbol{\Gamma}\mathbf{P}_{\boldsymbol{\theta}}^{\perp} - \mathbf{I}\right)\mathbf{z}(\ell), \tag{20}$$

where $\hat{\mathbf{H}}$ is the matrix of the previously estimated RTFs, $\mathbf{P}_{\boldsymbol{\theta}}^{\perp}$ is defined in (10c), and $\boldsymbol{\Gamma}$ is defined in (16c). The explicit expressions for the derivative in (20) w.r.t. the real and the imaginary part of θ_m are given by

$$\begin{aligned}\frac{\partial}{\partial r_m}\left[\boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H\right] &= \mathbf{i}_m\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H - \\ &- 2r_m\boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H + \boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\mathbf{i}_m^T, \end{aligned} \tag{21a}$$

$$\begin{aligned}\frac{\partial}{\partial c_m}\left[\boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H\right] &= j\mathbf{i}_m\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H - \\ &- 2c_m\boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^H - j\boldsymbol{\theta}\left(\boldsymbol{\theta}^H\boldsymbol{\theta}\right)^{-1}\mathbf{i}_m^T. \end{aligned} \tag{21b}$$

To complete the gradient derivation we substitute (20), (21a) and (21b) into (18) and simplify the expression using straightforward algebra, which results in (15) .

REFERENCES

[1] Alexey Ozerov and Cédric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[2] Hirokazu Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2009, pp. 3437–3440.

[3] Shoko Araki, Hiroshi Sawada, Ryo Mukai, and Shoji Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[4] Michael Syskind Pedersen, Jan Larsen, Ulrik Kjems, and Lucas C Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, pp. 1065–1094. New York, NY, USA: Springer, 2008.

[5] Barry D. Van Veen and Kevin M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[6] Simon Doclo and Marc Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[7] Tomohiro Nakatani, Nobutaka Ito, Takuya Higuchi, Shoko Araki, and Keisuke Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2017, pp. 286–290.

[8] Shlomo E. Chazan, Sharon Gannot, and Jacob Goldberger, "Attention-based neural network for joint diarization and speaker extraction," in *Proceedings of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[9] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, Alexey Ozerov, Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[10] Sharon Gannot, David Burshtein, and Ehud Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[11] Jingdong Chen, Jacob Benesty, and Yiteng Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 3, pp. 481–493, 2008.

[12] Ronen Talmon, Israel Cohen, and Sharon Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 546–555, 2009.

[13] Ernst Warsitz, Alexander Krueger, and Reinhold Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 73–76.

[14] Ofer Schwartz, Sharon Gannot, and Emanuël AP Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 2, pp. 240–251, 2015.

[15] Bracha Laufer, Ronen Talmon, and Sharon Gannot, "Relative transfer function modeling for supervised source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*, 2013, pp. 1–4.

[16] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[17] Ofir Shalvi and Ehud Weinstein, "System identification using nonstationary signals," *IEEE transactions on signal processing*, vol. 44, no. 8, pp. 2055–2063, 1996.

[18] Israel Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[19] Maja Taseska and Emanuël AP Habets, "Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1291–1304, 2016.

[20] Romain Serizel, Marc Moonen, Bas Van Dijk, and Jan Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, 2014.

[21] Ernst Warsitz and Reinhold Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[22] Shmulik Markovich-Golan and Sharon Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 544–548.

[23] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010*, pp. 201–204.

[24] Elior Hadad, Simon Doclo, and Sharon Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 543–558, 2016.

[25] Amin Hassani, Alexander Bertrand, and Marc Moonen, "LCMV beamforming with subspace projection for multi-speaker speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 91–95.

[26] Antoine. Deleforge, Sharon. Gannot, and Walter Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), Nice, France*, 2015.

[27] Shmulik Markovich-Golan, Sharon Gannot, and Walter Kellermann, "Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 320–332, 2017.

[28] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: a versatile framework for multichannel blind signal processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004, vol. 3, pp. 889–992.

[29] Yuanhang Zheng, Klaus Reindl, and Walter Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *The 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009, pp. 253–256.

[30] Selahattin Kayalar and Howard L Weinert, "Oblique projections: Formulas, algorithms, and error bounds," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 2, no. 1, pp. 33–45, 1989.

[31] Nicoleta Roman and DeLiang Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[32] Richard T. Behrens and Louis L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Transactions on Signal Processing*, vol. 42, no. 6, pp. 1413–1424, 1994.

[33] Gene H Golub and Charles F Van Loan, *Matrix computations*, vol. 3, JHU Press, 2012.

[34] Rémy Boyer, "Oblique projection for source estimation in a competitive environment: Algorithm and statistical analysis," *Signal Processing*, vol. 89, no. 12, pp. 2547–2554, 2009.

[35] Gal Reuven, Sharon Gannot, and Israel Cohen, "Dual-source transfer-function generalized sidelobe canceller," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 4, pp. 711–727, 2008.

[36] Dani Cherkassky, Shlomo E. Chazan, Jacob Goldberger, and Sharon Gannot, "Successive relative transfer function identification using single microphone speech enhancement," in *Proceedings of the 25th European Signal Processing Conference (EUSIPCO), Kos, Greece*, 2017.

[37] O. Walter, L. Drude, and R. Haeb-Umbach, "Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite gaussian mixture model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 459–463.

[38] Shlomo E Chazan, Jacob Goldberger, and Sharon Gannot, "DNN-based concurrent speakers detector and its application to speaker extraction with lcmv beamforming," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2018, pp. 6712–6716.

[39] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot, "Multichannel audio database in various acoustic environments," in *Proceedings of the 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.