Closed-Form Successive Relative Transfer Function Vector Estimation Based on Blind Oblique Projection Incorporating Noise Whitening

Henri Gode, *Graduate Student Member, IEEE*, and Simon Doclo, *Senior Member, IEEE*

***Abstract*—Relative transfer functions (RTFs) of sound sources play a crucial role in beamforming, enabling effective noise and interference suppression. This paper addresses the challenge of online estimating the relative transfer function (RTF) vectors of multiple sound sources in noisy and reverberant environments, both addressing scenarios with source activations and deactivations. To estimate the RTF vector of a newly activating source, the conventional blind oblique projection (BOP) method relies on computationally expensive gradient descent, random additional vectors, and high signal-to-noise ratio (SNR) conditions. We propose a generalized method that replaces iterative optimization with a closed-form solution, utilizes orthogonal additional vectors to improve accuracy, and integrates noise whitening and subtraction to ensure robustness in low SNR scenarios. Simulations are performed using real-world reverberant noisy recordings, featuring three successively activating speakers. In addition, simulations are performed for highly dynamic scenarios with random source activations and deactivations. Simulation results with and without a-priori knowledge of the source activity pattern demonstrate that the proposed BOP method with orthogonal additional vectors and noise whitening outperforms the conventional BOP method and other baseline methods in terms of computational efficiency and signal-to-interferer-and-noise ratio improvement, when applying the estimated RTF vectors in a linearly constrained minimum variance beamformer.**

***Index Terms*—Relative transfer function vectors, LCMV beamforming, Successive sources, Blind oblique projection, Covariance whitening.**

I. Introduction

In many hands-free speech communication applications, such as hearing aids, mobile phones, and smart speakers, interfering sounds and ambient noise may degrade the recorded microphone signals [1]. In such applications, online speech enhancement methods, which rely only on past information, are required to improve the quality and intelligibility of a target speaker. When multiple microphones are available, beamforming is a widely used technique to enhance a target speaker while suppressing interfering sources and noise [2], [3], [4]. Commonly used beamformers are the minimum variance distortionless response (MVDR) beamformer and the linearly constrained minimum variance (LCMV) beamformer [2]. Besides an estimate of the noise covariance matrix, these beamformers require an estimate of the RTF vector of the target source and possibly also the RTF vectors of the interfering sources. The RTF vector relates the acoustic transfer functions between a source and all microphones to a reference microphone, and plays an important role not only in beamforming, but also in, e.g., source localization [5], [6], [7], [8] and joint noise reduction and dereverberation [9], [10].

A. Related Works

Over the last decades, several methods have been proposed to estimate the RTF vector of a single source in a noisy environment, e.g., based on (weighted) least-squares [11], [12], [13], using maximum likelihood estimation [14], exploiting frequency correlations [15], using subspace decomposition methods such as covariance subtraction (CS) and covariance whitening (CW) [16], [17], [18], [19], [20], using manifold learning [21], [22], or deep neural network (DNN)-based mask estimation [23]. The CS method estimates the RTF vector by computing the principal eigenvector of the noisy covariance matrix after subtracting an estimate of the noise covariance matrix. On the other hand, the CW method estimates the RTF vector by de-whitening the principal eigenvector of the whitened noisy covariance matrix, where an estimate of the noise covariance matrix is used for the whitening operation. A performance comparison in [20] demonstrated that the CW method achieves superior results compared to the CS method.

In contrast to estimating the RTF vector of a single source, estimating the RTF vectors of multiple simultaneously active sources is more challenging. While the CS and CW methods can estimate the joint multi-source RTF subspace [17], and [24] introduces a generalized joint multi-source RTF representation using Plücker coordinates, these methods do not estimate the individual RTF vectors of each source. Joint diagonalization methods [25], [26] directly optimize the parameters of the underlying rank-1 signal model to estimate individual RTF vectors, but these methods fundamentally rely on post-processing to handle

Received 8 August 2025; revised 19 February 2026 and 16 April 2026; accepted 16 April 2026. Date of publication 23 April 2026; date of current version 15 May 2026. This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through Germany Excellence Strategy EXC 2177/2 - Project ID 390895286. The associate editor coordinating the review of this article and approving it for publication was Dr. Nobutaka Ito. *(Corresponding author: Henri Gode.)*

The authors are with the Department of Medical Physics and Acoustics and the Cluster of Excellence "Hearing4all.connects", Carl von Ossietzky Universität Oldenburg, 26129 Oldenburg, Germany (e-mail: henri.gode@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

Digital Object Identifier 10.1109/TASLPRO.2026.3686997

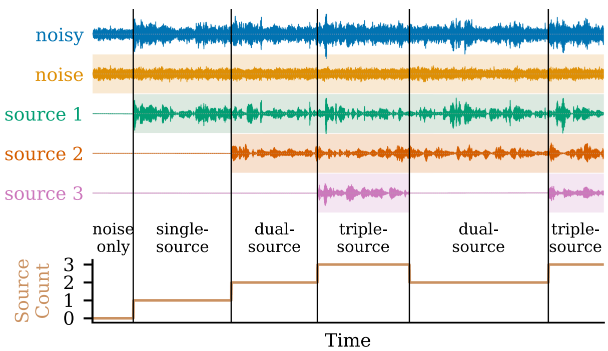

Fig. 1. Exemplary scenario with three activating and deactivating sources. Upper: Waveforms of one microphone signal and its speech and noise components for different time segments. Lower: Source count over time.

frequency permutation ambiguity. The recursive method in [27] solves an orthogonal Procrustes problem to estimate RTF vectors of all sources, but requires accurate initialization and known array geometry. In [28], individual RTF vectors are estimated by the TRINICON blind source separation framework [29], which however relies on geometric constraints based on direction-of-arrival (DOA) estimates and the array geometry. Alternatively, expectation maximization (EM) algorithms assuming W-disjoint orthogonality [30] can estimate RTF vectors of multiple sources, but they suffer from high computational complexity, challenging initialization, and frequency permutation ambiguity. In [31], an MVDR beamformer is guided by estimated covariance matrices and RTF vectors of an online complex Gaussian mixture model mask estimator, which however requires spatial priors based on the array geometry and source positions. Similarly, the guided source separation (GSS) method [32] employs a complex angular central Gaussian mixture model (cACGMM) mask estimator guided by diarization labels to compute covariance matrices for Souden MVDR beamforming [33]. Although the GSS method does not explicitly estimate RTF vectors, its online version [34] is included as a baseline for target speaker extraction in our experimental evaluation. Although the iterative source steering (ISS)-based independent vector analysis (IVA) method in [35] supports flexible online updates of individual source RTF vectors, it is not well-suited for the considered scenarios due to its noise-free rank-1 model and the assumption that the numbers of sources and microphones are equal.

In this paper, we consider online estimation of RTF vectors for successively activating and deactivating sources (exemplified in Fig. 1 featuring three sources). Although we will only consider scenarios involving source activations for the derivation of the methods, it will be shown that the resulting methods are also applicable to scenarios involving source deactivations. For successively activating sources, the RTF vector of the first source can be estimated using well-established single-source techniques like the CW method. To estimate the RTF vectors of subsequent sources (i.e., the second source in the dual-source segment and the third source in the triple-source segment), the covariance whitening with undesired covariance (CWu) method [20] can be employed as a baseline, which conceptually converts the multi-source problem into a single-source problem by treating all previously active sources as part of the background noise. Other approaches utilize estimates of the RTF vectors of already active sources to estimate the RTF vector of a newly activating source. For the dual-source case, the covariance blocking and whitening method in [36] estimates the RTF vector of the second source by solving a set of non-linear equations. The BOP method [37] estimates the RTF vector of a newly activating source by determining the oblique projection operator that optimally blocks this source while keeping the already active sources distortionless. However, the practical utility of the BOP method is limited by several issues: 1) the BOP method relies on an iterative gradient descent method, possibly converging to a local minimum and resulting in high computational complexity, 2) the BOP method employs random additional vectors, resulting in a large sensitivity to model mismatch, and 3) the BOP method is designed for high SNR, limiting its performance at low SNR. As for all successive RTF vector estimation methods, the BOP method requires knowledge about the source activity pattern, i.e., when sources become active and inactive. To determine the source activity pattern, several methods have been proposed for counting the number of active sources within a given signal segment, e.g., performing clustering on spatial features [38], [39], [40] or based on DNNs [41], [42]. More recently, several DNN-based approaches have been proposed that specifically address the problem of low-latency online source counting [43], [44], [45].

B. Contributions

To address the limitations of the conventional BOP method [37], in this paper we propose three extensions. First, we derive a closed-form solution to the BOP optimization problem, enhancing robustness and significantly reducing computational complexity. Second, we propose to utilize orthogonal additional vectors instead of random vectors, reducing estimation errors caused by model mismatch. Third, we incorporate noise handling techniques inspired by the CS and CW methods, making the BOP method suitable for low SNR conditions. For three successively activating sources in a reverberant noisy environment, simulation results for a binaural hearing aid setup with six microphones demonstrate that the proposed method featuring all three extensions outperforms the conventional BOP method for several source positions and SNRs. The improvements are evident with and without a-priori knowledge of the source activity pattern in terms of computational complexity, RTF vector estimation accuracy, and signal-to-interferer-and-noise ratio (SINR) improvement when the estimated RTF vectors are used in an LCMV beamformer. Furthermore, in a more realistic scenario with random source activations and deactivations, the proposed method also outperforms the conventional BOP method and the baseline CWu and GSS methods in terms of SINR improvement.

The remainder of this paper is organized as follows. Section II describes the signal model used for RTF vector estimation of successively activating sources. Section III introduces the conventional BOP method for successive RTF vector estimation. In Section IV we propose three extensions, offering a closed-form solution to the BOP optimization problem, introducing orthogonal additional vectors and incorporating noise handling. Section V compares the performance of the proposed method to the conventional BOP method and two other baseline methods in terms of RTF vector estimation accuracy and target source extraction performance.

II. Signal Model

We consider an acoustic scenario with $K_{\max}$ spatially stationary sound sources in a noisy and reverberant environment where the sound sources activate successively. The sound sources are recorded by an array with M microphones, with $K_{\max} \leq M$. The short-time Fourier transform (STFT) coefficients of the microphone signals at time frame t and frequency bin f are denoted by

$$\mathbf{y}_{t,f} = \begin{bmatrix} y_{1,t,f} & \cdots & y_{M,t,f} \end{bmatrix}^{\mathrm{T}} \in \mathbb{C}^{M\times 1}, \tag{1}$$

where $\{\cdot\}^{\mathrm{T}}$ denotes the transpose operator. The frequency index f will be omitted for notational brevity (i.e., $\mathbf{y}_t = \mathbf{y}_{t,f}$), unless explicitly required. At each time frame t, the multi-channel microphone signals $\mathbf{y}_t$ can be written as the sum of the source components $\mathbf{x}_{k,t}$ ($k \in \{1,\ldots,K_{\max}\}$) and the noise component $\mathbf{n}_t$, i.e.,

$$\mathbf{y}_t = \underbrace{\sum_{k=1}^{K_{\max}} \mathcal{I}_{k,t}\mathbf{x}_{k,t}}_{=\mathbf{x}_t} + \mathbf{n}_t, \tag{2}$$

where the vectors $\mathbf{x}_{k,t} \in \mathbb{C}^{M\times 1}$ and $\mathbf{n}_t \in \mathbb{C}^{M\times 1}$ are defined similarly as in (1) and $\mathcal{I}_{k,t} \in \{0,1\}$ denotes the activity of the k-th source, which is not assumed to be frequency-dependent. In this paper, we consider successively activating sources, i.e.,

$$\mathcal{I}_{k,t} = \begin{cases} 0, & \text{if } t < t_k \\ 1, & \text{if } t \geq t_k \end{cases}, \tag{3}$$

where t_k denotes the activation time of the k-th source. We assume that simultaneous activation of multiple sources does not occur, i.e., there is at least one frame between the activation of two sources ($t_{k+1} - t_k > 0$). Assuming the source components and the noise component are uncorrelated, the noisy covariance matrix $\mathbf{R}_{y,t} = \mathbb{E}\{\mathbf{y}_t\mathbf{y}_t^{\mathrm{H}}\}$, with $\mathbb{E}\{\cdot\}$ denoting the expectation operator and $\{\cdot\}^{\mathrm{H}}$ denoting the conjugate transpose operator, can be written as

$$\mathbf{R}_{y,t} = \mathbf{R}_{x,t} + \mathbf{R}_n, \tag{4}$$

with $\mathbf{R}_{x,t} = \mathbb{E}\{\mathbf{x}_t\mathbf{x}_t^{\mathrm{H}}\}$ the noiseless covariance matrix and $\mathbf{R}_n = \mathbb{E}\{\mathbf{n}_t\mathbf{n}_t^{\mathrm{H}}\}$ the noise covariance matrix, which is assumed to be time-invariant and full-rank.

Assuming sufficiently large STFT frames, the k-th source component $\mathbf{x}_{k,t}$ can be modeled as the multiplication of the k-th source component in a reference microphone, denoted by r, with the (time-invariant) RTF vector $\mathbf{g}_k \in \mathbb{C}^{M\times 1}$ of the k-th source [46], i.e.,

$$\mathbf{x}_{k,t} = \mathbf{g}_k x_{k,r,t} \quad \forall k \in \{1,\ldots,K_{\max}\}. \tag{5}$$

It should be noted that the reference entry of all RTF vectors equals 1, i.e., $\mathbf{e}_r^{\mathrm{T}}\mathbf{g}_k = 1$, where the r-th entry (corresponding to the reference microphone) of the selection vector $\mathbf{e}_r$ is equal to 1 and all other entries are equal to 0. We assume that the RTF vectors of all sources are linearly independent. The problem of estimating the RTF vectors of all sources can be reformulated into the problem of successively estimating the RTF vector of each source in its activating time segment $\mathcal{T}_k = [t_k, t_{k+1}[$. Considering the K-th time segment, the corresponding subproblem can be seen as a scenario with $K \leq K_{\max}$ active sources, where the K-th source denotes the newest activating source, i.e.,

$$\mathbf{y}_t = \underbrace{\mathbf{x}_{K,t}}_{\text{new}} + \underbrace{\sum_{k=1}^{K-1}\mathbf{x}_{k,t}}_{\text{old}} + \mathbf{n}_t = \mathbf{g}_K x_{K,r,t} + \underbrace{\sum_{k=1}^{K-1}\mathbf{g}_k x_{k,r,t} + \mathbf{n}_t}_{=\mathbf{v}_t}, \tag{6}$$

where $\mathbf{v}_t \in \mathbb{C}^{M\times 1}$ denotes the undesired component. By stacking the RTF vectors of the already active sources in the matrix

$$\mathbf{G}_{\bar{K}} = \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 & \cdots & \mathbf{g}_{K-1} \end{bmatrix} \in \mathbb{C}^{M\times(K-1)}, \tag{7}$$

the noiseless covariance matrix can be written as

$$\boxed{\mathbf{R}_{x,t} = \underbrace{\mathbf{g}_K \phi_{K,t}\mathbf{g}_K^{\mathrm{H}}}_{=\mathbf{R}_{K,t}} + \underbrace{\mathbf{G}_{\bar{K}}\mathbf{\Phi}_{\bar{K},t}\mathbf{G}_{\bar{K}}^{\mathrm{H}}}_{=\mathbf{R}_{\bar{K},t}}} \tag{8}$$

The matrices $\mathbf{R}_{K,t}$ and $\mathbf{R}_{\bar{K},t}$ denote the $M \times M$-dimensional covariance matrices of the K-th activating (new) source and the $K-1$ already active (old) sources, which have rank 1 and rank $K-1$, respectively. The diagonal matrix $\mathbf{\Phi}_{\bar{K},t} = \operatorname{diag}\{[\phi_{1,t} \quad \phi_{2,t} \quad \cdots \quad \phi_{K-1,t}]\} \in \mathbb{R}_+^{(K-1)\times(K-1)}$ contains the power spectral densities (PSDs) of the already active sources, with $\operatorname{diag}\{\cdot\}$ building a diagonal matrix from a vector and $\phi_{k,t} = \mathbb{E}\{|x_{k,r,t}|^2\}$ denoting the (time-varying) power spectral densities (PSDs) of the k-th source component in the reference microphone.

In practice, the noisy covariance matrix $\mathbf{R}_{y,t}$ can be estimated recursively using an exponential sliding window, i.e.,

$$\widehat{\mathbf{R}}_{y,t} = \alpha\widehat{\mathbf{R}}_{y,t-1} + (1-\alpha)\,\mathbf{y}_t\mathbf{y}_t^{\mathrm{H}}, \tag{9}$$

where $\alpha = e^{-\frac{t_{\mathrm{fs}}}{t_\alpha}}$ denotes the forgetting factor, with t_α and t_{fs} the smoothing time constant and the STFT frame shift, respectively. This paper focuses on online estimating the RTF vector $\mathbf{g}_K$ of the newly activating source, assuming that estimates of the noise covariance matrix $\mathbf{R}_n$ and the RTF vectors $\mathbf{G}_{\bar{K}}$ of the already active sources are available. The RTF vectors $\mathbf{g}_1 \ldots \mathbf{g}_{K-1}$ can be successively estimated in the first $(K-1)$ time segments, whereas the noise covariance matrix can be estimated in a noise-only segment. We will first assume that the source activity pattern $\mathcal{I}_{k,t}$ is known, while in Section V we will evaluate the performance of all successive RTF vector estimation methods both using oracle knowledge of the source activity pattern as well as using the online source counting method from [45] to estimate the source activity pattern.

III. Conventional BOP Method

In this section, we will review the conventional BOP method from [37], which serves as the basis for the extensions proposed in Section IV. To estimate the RTF vector $\mathbf{g}_K$ of the source activating in the K-th segment, the conventional BOP method aims at blocking this source while keeping the $K-1$ already active sources distortionless. The BOP method utilizes the oblique projection operator $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}$, i.e.,

$$\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} = \mathbf{G}_{\bar{K}}\left(\mathbf{G}^{\mathrm{H}}_{\bar{K}}\mathbf{P}^{\perp}_{\boldsymbol{\theta}}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}^{\mathrm{H}}_{\bar{K}}\mathbf{P}^{\perp}_{\boldsymbol{\theta}}, \tag{10}$$

$$\text{with} \quad \mathbf{P}^{\perp}_{\boldsymbol{\theta}} = \mathbf{I}_M - \boldsymbol{\theta}\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^{\mathrm{H}}, \tag{11}$$

with $\mathbf{G}_{\bar{K}}$ defined in (7) containing the known RTF vectors of the already active sources and $\boldsymbol{\theta} \in \mathbb{C}^{M\times 1}$ a vector variable. The matrix $\mathbf{P}^{\perp}_{\boldsymbol{\theta}}$ in (11) denotes the complement orthogonal projection operator, which projects vectors onto the subspace orthogonal to the vector $\boldsymbol{\theta}$, and thus blocks the vector $\boldsymbol{\theta}$, i.e.,

$$\mathbf{P}^{\perp}_{\boldsymbol{\theta}}\boldsymbol{\theta} = \mathbf{0}_{M\times 1}. \tag{12}$$

where $\mathbf{0}_{M\times N}$ denotes the $M \times N$-dimensional zero matrix. By defining $[\mathbf{G}_{\bar{K}} \quad \boldsymbol{\theta}]_{\perp}$ as the orthogonal complement of the matrix $[\mathbf{G}_{\bar{K}} \quad \boldsymbol{\theta}]$, i.e., the matrix containing $(M-K)$ orthonormal vectors spanning the subspace orthogonal to the column space of $[\mathbf{G}_{\bar{K}} \quad \boldsymbol{\theta}]$, and using $\mathbf{P}^{\perp}_{\boldsymbol{\theta}}[\mathbf{G}_{\bar{K}} \quad \boldsymbol{\theta}]_{\perp} = [\mathbf{G}_{\bar{K}} \quad \boldsymbol{\theta}]_{\perp}$, and $\mathbf{G}^{\mathrm{H}}_{\bar{K}}[\mathbf{G}_{\bar{K}} \quad \boldsymbol{\theta}]_{\perp} = \mathbf{0}_{(K-1)\times(M-K)}$, the following properties of the oblique projection operator $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}$ in (10) can be derived:

$$\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{G}_{\bar{K}} = \mathbf{G}_{\bar{K}}, \quad \mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\boldsymbol{\theta} = \mathbf{0}_{M\times 1}, \tag{13}$$

$$\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\begin{bmatrix}\mathbf{G}_{\bar{K}} & \boldsymbol{\theta}\end{bmatrix}_{\perp} = \mathbf{0}_{M\times(M-K)}. \tag{14}$$

Aiming at blocking the K-th source while keeping the $K-1$ already active sources distortionless using the oblique projection operator $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}$ in (10), the BOP cost function is defined as the power of the projected signal, i.e.,

$$J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}) = \mathrm{tr}\left\{\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_{y,t}\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right\}, \tag{15}$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace. Using the signal model in (8) and the fact that $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{G}_{\bar{K}} = \mathbf{G}_{\bar{K}}$ (which corresponds to keeping the already active sources distortionless), the BOP cost function in (15) can be written as

$$\begin{aligned} J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}) &= \mathrm{tr}\left\{\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{g}_K\phi_{K,t}\mathbf{g}^{\mathrm{H}}_K\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right\} \\ &+ \mathrm{tr}\left\{\mathbf{G}_{\bar{K}}\mathbf{\Phi}_{\bar{K},t}\mathbf{G}^{\mathrm{H}}_{\bar{K}}\right\} + \mathrm{tr}\left\{\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_n\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right\}, \end{aligned} \tag{16}$$

where the term $\mathrm{tr}\{\mathbf{G}_{\bar{K}}\mathbf{\Phi}_{\bar{K},t}\mathbf{G}^{\mathrm{H}}_{\bar{K}}\}$ is a positive constant (independent of $\boldsymbol{\theta}$), and both other terms depend on the variable $\boldsymbol{\theta}$. The conventional BOP method in [37] assumes a sufficiently large SNR, such that the noise covariance matrix $\mathbf{R}_n$ can be neglected in (4) and the noisy covariance matrix $\mathbf{R}_{y,t}$ reduces to the noiseless covariance matrix $\mathbf{R}_{x,t}$. Under this assumption, the BOP cost function is equal to

$$J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}) = \mathrm{tr}\left\{\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_{x,t}\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right\} \tag{17}$$

$$= \mathrm{tr}\left\{\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{g}_K\phi_{K,t}\mathbf{g}^{\mathrm{H}}_K\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right\} + \mathrm{tr}\left\{\mathbf{G}_{\bar{K}}\mathbf{\Phi}_{\bar{K},t}\mathbf{G}^{\mathrm{H}}_{\bar{K}}\right\}, \tag{18}$$

where the only term depending on $\boldsymbol{\theta}$ is the first term $\mathrm{tr}\{\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{g}_K\phi_{K,t}\mathbf{g}^{\mathrm{H}}_K\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\}$, involving the RTF vector $\mathbf{g}_K$ of the K-th source, which is to be estimated. The cost function in (18) is minimized when $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{g}_K = \mathbf{0}_{M\times 1}$, i.e., when $\boldsymbol{\theta}$ is equal to $\mathbf{g}_K$ (or a scaled version), so that the K-th source is blocked by the oblique projection operator $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}$ and the first term in (18) is zero. Therefore, an estimate of the RTF vector $\mathbf{g}_K$ of the K-th source can be obtained as

$$\boxed{\widehat{\mathbf{g}}_K = \frac{\tilde{\mathbf{g}}_K}{\mathbf{e}^{\mathrm{T}}_r\tilde{\mathbf{g}}_K} \quad \text{with} \quad \tilde{\mathbf{g}}_K = \arg\min_{\boldsymbol{\theta}}\left(J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta})\right)} \tag{19}$$

Since no closed-form solution for the optimization problem in (19) exists, it was proposed in [37] to use an iterative gradient descent method with an element-wise gradient.

IV. Proposed Extensions

In this section, we propose three extensions to the conventional BOP method, which significantly reduce computational complexity and enhance robustness. After deriving the vector gradient of the BOP cost function and reviewing the use of additional random vectors in the conventional BOP method, we derive a closed-form solution to the BOP cost function in Section IV-A. In Section IV-B, we suggest a more appropriate choice of additional vectors instead of random vectors. In Section IV-C, we generalize the BOP method to low SNR conditions by integrating noise handling, motivated by the covariance subtraction (CS) and covariance whitening (CW) methods. Although the derivations in Sections IV-A and IV-B are based on the noiseless covariance matrix (i.e., assuming no noise is present), the conclusions derived in these sections remain valid when noise is present if the noise handling proposed in Section IV-C is used.

A. Closed-Form Solution

In [37] the element-wise gradient of the BOP cost function in (17) was derived. Instead of the element-wise gradient, here we will consider the vector gradient since it has a lower computational complexity and is more convenient for later derivations. The vector gradient of the BOP cost function in (17) is equal to (see detailed derivation in Appendix A):

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}) &= -\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\mathbf{G}_{\bar{K}}}\boldsymbol{\theta}\left(\boldsymbol{\theta}^{\mathrm{H}}\mathbf{P}^{\perp}_{\mathbf{G}_{\bar{K}}}\boldsymbol{\theta}\right)^{-1} \\ &-\mathbf{P}^{\perp}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}\mathbf{R}_{x,t}\mathbf{P}^{\angle\mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{G}_{\bar{K}}\left(\mathbf{G}^{\mathrm{H}}_{\bar{K}}\mathbf{P}^{\perp}_{\boldsymbol{\theta}}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}^{\mathrm{H}}_{\bar{K}}\boldsymbol{\theta}\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)^{-1}. \end{aligned} \tag{20}$$

where the matrices $\mathbf{P}^{\perp}_{\mathbf{G}_{\bar{K}}}$ and $\mathbf{P}^{\perp}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}$ denote the complement orthogonal projection operators defined as in (11), projecting onto the subspace orthogonal to the column space of $\mathbf{G}_{\bar{K}}$ and the column space of $[\mathbf{G}_{\bar{K}} \quad \boldsymbol{\theta}]$, respectively. Following (12), they also satisfy

$$\mathbf{P}^{\perp}_{\mathbf{G}_{\bar{K}}}\mathbf{G}_{\bar{K}} = \mathbf{0}_{M\times(K-1)}, \quad \mathbf{P}^{\perp}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}\begin{bmatrix}\mathbf{G}_{\bar{K}} & \boldsymbol{\theta}\end{bmatrix} = \mathbf{0}_{M\times K}. \tag{21}$$

Using the signal model in (8) and the properties in (13) and (21), it follows that for $\boldsymbol{\theta} = \mathbf{g}_K$, both $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\mathbf{G}_{\bar{K}}}$ and

$\mathbf{P}^{\perp}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}\mathbf{R}_{x,t}$ are equal to $\mathbf{0}_{M\times M}$. Consequently, it can be easily verified that for $\boldsymbol{\theta}=\mathbf{g}_K$ the gradient in (20) is equal to $\mathbf{0}_{M\times 1}$, which is consistent with the fact that $\boldsymbol{\theta}=\mathbf{g}_K$ minimizes the BOP cost function in (17). However, when $K<M$, the null space of the rank-K matrix $[\mathbf{g}_K \;\; \mathbf{G}_{\bar{K}}]^{\mathrm{H}}\in\mathbb{C}^{K\times M}$ is non-empty and consists of vectors $\tilde{\boldsymbol{\theta}}$ for which $\mathbf{g}_K^{\mathrm{H}}\tilde{\boldsymbol{\theta}}=0$ and $\mathbf{G}_{\bar{K}}^{\mathrm{H}}\tilde{\boldsymbol{\theta}}=\mathbf{0}_{(K-1)\times 1}$, i.e., $\tilde{\boldsymbol{\theta}}\perp\mathbf{G}_{\bar{K}}$. Using (10)–(14) and defining the orthogonal projection operator $\mathbf{P}^{\parallel}_{\mathbf{G}_{\bar{K}}}=\mathbf{G}_{\bar{K}}(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{G}_{\bar{K}})^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}$ with the property $\mathbf{P}^{\parallel}_{\mathbf{G}_{\bar{K}}}\mathbf{G}_{\bar{K}}=\mathbf{G}_{\bar{K}}$, it can be shown that the oblique projection operator $\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}$ reduces to the orthogonal projection operator $\mathbf{P}^{\parallel}_{\mathbf{G}_{\bar{K}}}$ for all vectors $\tilde{\boldsymbol{\theta}}\perp\mathbf{G}_{\bar{K}}$, i.e.,

$$\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\tilde{\boldsymbol{\theta}}}=\mathbf{P}^{\parallel}_{\mathbf{G}_{\bar{K}}}. \tag{22}$$

Hence, it follows that the BOP cost function in (18) for all vectors $\tilde{\boldsymbol{\theta}}$ is equal to

$$J_{\mathbf{G}_{\bar{K}}}\left(\tilde{\boldsymbol{\theta}}\right)=\operatorname{tr}\left\{\mathbf{P}^{\parallel}_{\mathbf{G}_{\bar{K}}}\mathbf{g}_K\phi_{K,t}\mathbf{g}_K^{\mathrm{H}}\mathbf{P}^{\parallel\mathrm{H}}_{\mathbf{G}_{\bar{K}}}\right\}+\operatorname{tr}\left\{\mathbf{G}_{\bar{K}}\boldsymbol{\Phi}_{\bar{K},t}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\right\}. \tag{23}$$

Since the cost function in (23) does not depend on $\boldsymbol{\theta}$, for all vectors $\tilde{\boldsymbol{\theta}}$ in the null space of $[\mathbf{g}_K \;\; \mathbf{G}_{\bar{K}}]^{\mathrm{H}}$ the gradient is equal to zero, revealing a (potentially multidimensional) plateau of the BOP cost function and posing a challenge for gradient descent methods. To overcome this problem, it was proposed in [37] to add $M-K$ random vectors $\mathbf{G}_a=[\mathbf{g}_{K+1} \;\; \cdots \;\; \mathbf{g}_M]$ to the RTF vectors of the already active sources, i.e., $\tilde{\mathbf{G}}_{\bar{K}}=[\mathbf{G}_{\bar{K}} \;\; \mathbf{G}_a]$ with $\operatorname{rank}\{\tilde{\mathbf{G}}_{\bar{K}}\}=M-1$, so that the matrix $[\mathbf{g}_K \;\; \tilde{\mathbf{G}}_{\bar{K}}]$ is full rank and its null space is empty (assuming no random vector lies in the column space of $[\mathbf{g}_K \;\; \mathbf{G}_{\bar{K}}]$). Note that introducing these additional random vectors changes the BOP cost function in (17) and the gradient in (20) to

$$J_{\tilde{\mathbf{G}}_{\bar{K}}}(\boldsymbol{\theta})=J_{[\mathbf{G}_{\bar{K}},\mathbf{G}_a]}(\boldsymbol{\theta})=\operatorname{tr}\left\{\mathbf{P}^{\angle}_{\tilde{\mathbf{G}}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_{x,t}\mathbf{P}^{\angle\mathrm{H}}_{\tilde{\mathbf{G}}_{\bar{K}}\boldsymbol{\theta}}\right\}, \tag{24}$$

$$\nabla_{\boldsymbol{\theta}}J_{\tilde{\mathbf{G}}_{\bar{K}}}(\boldsymbol{\theta})=-\mathbf{P}^{\angle\mathrm{H}}_{\tilde{\mathbf{G}}_{\bar{K}}\boldsymbol{\theta}}\mathbf{P}^{\angle}_{\tilde{\mathbf{G}}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\boldsymbol{\theta}\left(\boldsymbol{\theta}^{\mathrm{H}}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\boldsymbol{\theta}\right)^{-1}, \tag{25}$$

where the second term of the gradient vanishes since $\operatorname{rank}\{[\tilde{\mathbf{G}}_{\bar{K}} \;\; \boldsymbol{\theta}]\}=M$ (except for $\boldsymbol{\theta}$ in the column space of $\tilde{\mathbf{G}}_{\bar{K}}$, which is highly improbable), so that $\mathbf{P}^{\perp}_{[\tilde{\mathbf{G}}_{\bar{K}},\boldsymbol{\theta}]}=\mathbf{0}_{M\times M}$. Using the signal model in (8) and the fact that for $\boldsymbol{\theta}=\mathbf{g}_K$ it follows that $\mathbf{P}^{\angle}_{\tilde{\mathbf{G}}_{\bar{K}}\boldsymbol{\theta}}\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}=\mathbf{0}_{M\times M}$, it can be shown that for $\boldsymbol{\theta}=\mathbf{g}_K$ the gradient in (25) is equal to $\mathbf{0}_{M\times 1}$, so that the cost function with the additional vectors $\mathbf{G}_a$ in (24) has the same global minimum as the original cost function in (17).

Using the vector gradient in (25), we now derive a closed-form solution minimizing the BOP cost function with additional vectors in (24). Let us consider the eigenvalue decomposition of the matrix $\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\in\mathbb{C}^{M\times M}$, i.e.,

$$\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\bar{\boldsymbol{\theta}}_m=\lambda_m\bar{\boldsymbol{\theta}}_m \quad \forall m\in\{1,\ldots,M\}, \tag{26}$$

where $\bar{\boldsymbol{\theta}}_m$ and λ_m denote the eigenvectors and eigenvalues, respectively. Since the rank of $\tilde{\mathbf{G}}_{\bar{K}}$ is equal to $M-1$, it follows from (11) and (21) that the rank of $\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}$ and consequently the rank of $\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}$ is equal to 1. Hence, all eigenvalues except for the principal eigenvalue λ_1 are equal to zero, i.e.,

$$\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\bar{\boldsymbol{\theta}}_1=\lambda_1\bar{\boldsymbol{\theta}}_1, \tag{27}$$

$$\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\bar{\boldsymbol{\theta}}_m=\mathbf{0}_{M\times 1} \quad \forall m\in\{2,\ldots,M\}, \tag{28}$$

where $\bar{\boldsymbol{\theta}}_1$ denotes the principal eigenvector. Using (27), it can be seen that the principal eigenvector sets the gradient in (25) to zero, i.e.,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}J_{\tilde{\mathbf{G}}_{\bar{K}}}(\bar{\boldsymbol{\theta}}_1)&=-\mathbf{P}^{\angle\mathrm{H}}_{\tilde{\mathbf{G}}_{\bar{K}}\bar{\boldsymbol{\theta}}_1}\mathbf{P}^{\angle}_{\tilde{\mathbf{G}}_{\bar{K}}\bar{\boldsymbol{\theta}}_1}\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\bar{\boldsymbol{\theta}}_1\left(\bar{\boldsymbol{\theta}}_1^{\mathrm{H}}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\bar{\boldsymbol{\theta}}_1\right)^{-1}\\&=-\mathbf{P}^{\angle\mathrm{H}}_{\tilde{\mathbf{G}}_{\bar{K}}\bar{\boldsymbol{\theta}}_1}\underbrace{\mathbf{P}^{\angle}_{\tilde{\mathbf{G}}_{\bar{K}}\bar{\boldsymbol{\theta}}_1}\bar{\boldsymbol{\theta}}_1}_{=\mathbf{0}_{M\times 1}}\lambda_1\left(\bar{\boldsymbol{\theta}}_1^{\mathrm{H}}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\bar{\boldsymbol{\theta}}_1\right)^{-1}.\end{aligned} \tag{29}$$

Since using the rank-nullity theorem the dimension of the null spaces of $\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}$ and $\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}$ are equal to $M-1$, their null spaces are the same, so that for all other eigenvectors $\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\bar{\boldsymbol{\theta}}_m=\mathbf{0}_{M\times 1}$ ($\forall m\in\{2,\ldots,M\}$). Hence, these eigenvectors cannot be solutions as this would cause a division by zero in the term $(\boldsymbol{\theta}^{\mathrm{H}}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\boldsymbol{\theta})^{-1}$ of the gradient in (25). The closed-form solution minimizing the BOP cost function $J_{\tilde{\mathbf{G}}_{\bar{K}}}(\boldsymbol{\theta})$ in (24) is hence given by

$$\boxed{\widehat{\mathbf{g}}_K=\frac{\tilde{\mathbf{g}}_K}{\mathbf{e}_r^{\mathrm{T}}\tilde{\mathbf{g}}_K}\quad\text{with}\quad\tilde{\mathbf{g}}_K=\mathcal{V}_{\max}\left\{\mathbf{R}_{x,t}\mathbf{P}^{\perp}_{\tilde{\mathbf{G}}_{\bar{K}}}\right\}} \tag{30}$$

with $\mathcal{V}_{\max}\{\cdot\}$ denoting the principal eigenvector.

B. Orthogonal Additional Vectors

As shown in Section IV-A, when the signal model in (8) holds, the choice of the $M-K$ additional vectors $\mathbf{G}_a$ has no influence on the global minimum of the BOP cost function. However, since in practice the signal model in (8) and its corresponding assumptions, e.g., the rank-1 approximation in (5) for each source, may not perfectly hold, it should be realized that the solution in (30) depends on the choice of additional vectors $\mathbf{G}_a$. It should be noted that the oblique projection operator $\mathbf{P}^{\angle}_{\tilde{\mathbf{G}}_{\bar{K}}\boldsymbol{\theta}}$ may amplify a vector when the angle between the column spaces of $\tilde{\mathbf{G}}_{\bar{K}}$ and $\boldsymbol{\theta}$ is small. This means that applying the oblique projection operator $\mathbf{P}^{\angle}_{\tilde{\mathbf{G}}_{\bar{K}}\boldsymbol{\theta}}$ in (25) for $\boldsymbol{\theta}=\mathbf{g}_K$ could potentially lead to amplification of error signal components caused by model mismatch when the angle between the column spaces of $\tilde{\mathbf{G}}_{\bar{K}}=[\mathbf{G}_{\bar{K}} \;\; \mathbf{G}_{\mathbf{a}}]$ and $\mathbf{g}_K$ is small. To reduce such amplification, a good choice of additional vectors $\mathbf{G}_a$ would hence be vectors that are orthogonal to $\mathbf{g}_K$. Since obviously $\mathbf{g}_K$ is not known, we propose to compute the $M-K$ additional vectors as the minor subspace of the covariance matrix $\mathbf{R}_{x,t}$, which are orthogonal to $\mathbf{g}_K$ assuming the signal model in (8) holds, i.e.,

$$\mathbf{G}_a=\mathcal{V}_{\min}^{M-K}\left\{\mathbf{R}_{x,t}\right\}, \tag{31}$$

with $\mathcal{V}_{\min}^{N}\{\cdot\}$ denoting the minor subspace of dimension N spanned by the eigenvectors corresponding to the N smallest eigenvalues. We call this method BOP using orthogonal additional vectors (BOPO).

C. Noise Handling

Similarly as in [37], in Sections IV-A and IV-B it has been assumed that no noise is present, i.e., all expressions have been derived using the noiseless covariance matrix $\mathbf{R}_{x,t}$. When noise is present, an additional noise term appears in the cost function in (24) similar to the last term in (16), which potentially shifts the global minimum of the cost function away from the true RTF vector $\mathbf{g}_K$ and hence introduces estimation errors in (30). To mitigate this effect, we aim to estimate the noiseless covariance matrix $\mathbf{R}_{x,t}$ from the noisy covariance matrix $\mathbf{R}_{y,t}$. Aiming at generalizing the BOP method to noisy scenarios, in this section we propose two noise handling techniques, which are motivated by the CS and CW methods for single-source RTF vector estimation [16], [17], [18], [19], [20]. Both techniques assume an estimate of the noise covariance matrix $\mathbf{R}_n$ to be available (e.g., estimated from a noise-only segment), which is used to construct a noiseless covariance matrix similar to $\mathbf{R}_{x,t}$. The conventional BOP method from [37] assumes that no noise is present, i.e., $\mathbf{R}_{x,t} = \mathbf{R}_{y,t}$.

1) Subtract the Noise Covariance Matrix: Based on (4), the noiseless covariance matrix $\mathbf{R}_{x,t}$ can be simply computed by subtracting the noise covariance matrix $\mathbf{R}_n$ from the noisy covariance matrix $\mathbf{R}_{y,t}$, i.e.,

$$\mathbf{R}_{x,t}^{(\mathrm{s})} = \mathbf{R}_{y,t} - \mathbf{R}_n. \tag{32}$$

The corresponding methods are called BOP with noise subtraction (BOP-S) and BOPO with noise subtraction (BOPO-S), respectively.

2) Whiten the Noise Covariance Matrix: Similarly as for the CW method [20], the noise covariance matrix $\mathbf{R}_n$ can also be used to compute the whitened noiseless covariance matrix as

$$\mathbf{R}_{x,t}^{(\mathrm{w})} = \mathbf{R}_n^{\frac{-\mathrm{H}}{2}} \mathbf{R}_{y,t} \mathbf{R}_n^{\frac{-1}{2}} - \mathbf{I}_M, \tag{33}$$

where $\mathbf{R}_n^{\frac{1}{2}}$ denotes a matrix square root decomposition (e.g., Cholesky decomposition) of the noise covariance matrix, i.e., $\mathbf{R}_n = \mathbf{R}_n^{\frac{\mathrm{H}}{2}} \mathbf{R}_n^{\frac{1}{2}}$. Spatially whitening the noise term and subtracting an identity matrix effectively removes the additional noise term from the cost function in (24). This can be interpreted as transforming the noisy covariance matrix $\mathbf{R}_{y,t}$ into a whitened domain to estimate the noiseless covariance matrix $\mathbf{R}_{x,t}^{(\mathrm{w})}$ in this whitened domain, allowing to apply the same methods (closed-form solution and orthogonal additional vectors) as derived for the original domain. However, before adding additional vectors $\mathbf{G}_a^{(\mathrm{w})}$, the RTF vectors of the already active sources also need to be transferred into the whitened domain, i.e.,

$$\mathbf{G}_{\bar{K}}^{(\mathrm{w})} = \mathbf{R}_n^{\frac{-\mathrm{H}}{2}} \mathbf{G}_{\bar{K}}, \tag{34}$$

such that $\tilde{\mathbf{G}}_{\bar{K}}^{(\mathrm{w})} = [\mathbf{G}_{\bar{K}}^{(\mathrm{w})} \quad \mathbf{G}_a^{(\mathrm{w})}]$. In addition, after using (30) to compute $\tilde{\mathbf{g}}_K^{(\mathrm{w})} = \mathcal{V}_{\max}\{\mathbf{R}_{x,t}^{(\mathrm{w})} \mathbf{P}_{\tilde{\mathbf{G}}_{\bar{K}}^{(\mathrm{w})}}\}$ in the whitened domain, this vector still needs to be transferred into the original domain by de-whitening, i.e.,

$$\tilde{\mathbf{g}}_K = \mathbf{R}_n^{\frac{\mathrm{H}}{2}} \tilde{\mathbf{g}}_K^{(\mathrm{w})}. \tag{35}$$

The reference entry normalization is then performed in the original domain as shown in (30). The corresponding methods are called BOP with noise whitening (BOP-W) and BOPO with noise whitening (BOPO-W), respectively. An overview of the proposed BOP-based RTF vector estimation method with the different noise handling techniques and additional vector choices is provided in Algorithm 1.

Algorithm 1: Proposed BOP-Based RTF Vector Estimation.

Require: Noisy covariance matrix $\mathbf{R}_{y,t}$, noise covariance matrix $\mathbf{R}_n$, RTF vectors of already active sources $\mathbf{G}_{\bar{K}}$

Noise Handling:
- **if** Subtraction (BOP-S, BOPO-S):
  - $\widehat{\mathbf{R}}_{x,t} = \mathbf{R}_{y,t} - \mathbf{R}_n$ by (32)
- **else if** Whitening (BOP-W, BOPO-W):
  - Compute $\mathbf{R}_n^{\frac{1}{2}}$
  - $\widehat{\mathbf{R}}_{x,t} = \mathbf{R}_n^{\frac{-\mathrm{H}}{2}} \mathbf{R}_{y,t} \mathbf{R}_n^{\frac{-1}{2}} - \mathbf{I}_M$ by (33)
  - $\mathbf{G}_{\bar{K}} = \mathbf{R}_n^{\frac{-\mathrm{H}}{2}} \mathbf{G}_{\bar{K}}$ by (34)
- **else** None (BOP, BOPO):
  - $\widehat{\mathbf{R}}_{x,t} = \mathbf{R}_{y,t}$

Additional Vectors:
- **if** Orthogonal (BOPO, BOPO-S, BOPO-W):
  - $\mathbf{G}_a = \mathcal{V}_{\min}^{M-K}\{\widehat{\mathbf{R}}_{x,t}\}$ by (31)
- **else** Random (BOP, BOP-S, BOP-W):
  - $\mathbf{G}_a =$ random vectors
- Construct $\tilde{\mathbf{G}}_{\bar{K}} = [\mathbf{G}_{\bar{K}} \quad \mathbf{G}_a]$

Closed-form RTF Estimation:
- $\mathbf{P}_{\tilde{\mathbf{G}}_{\bar{K}}}^{\perp} = \mathbf{I}_M - \tilde{\mathbf{G}}_{\bar{K}} (\tilde{\mathbf{G}}_{\bar{K}}^{\mathrm{H}} \tilde{\mathbf{G}}_{\bar{K}})^{-1} \tilde{\mathbf{G}}_{\bar{K}}^{\mathrm{H}}$
- $\tilde{\mathbf{g}}_K = \mathcal{V}_{\max}\{\widehat{\mathbf{R}}_{x,t} \mathbf{P}_{\tilde{\mathbf{G}}_{\bar{K}}}^{\perp}\}$ by (30)
- **if** Whitening (BOP-W, BOPO-W):
  - $\tilde{\mathbf{g}}_K = \mathbf{R}_n^{\frac{\mathrm{H}}{2}} \tilde{\mathbf{g}}_K$ by (35)
- $\widehat{\mathbf{g}}_K = \frac{\tilde{\mathbf{g}}_K}{\mathbf{e}_r^{\mathrm{T}} \tilde{\mathbf{g}}_K}$ by (30)
- **return:** Estimated RTF vector $\widehat{\mathbf{g}}_K$

V. Evaluation

In this section, we compare the performance of the proposed successive RTF vector estimation method using the extensions discussed in Section IV to the conventional BOP method. After introducing the acoustic scenario with three successively activating sources in Section V-A, algorithmic implementation details and computational complexity are discussed in Section V-B. Section V-C discusses the considered performance metrics, namely RTF vector estimation accuracy and SINR improvement of an LCMV beamformer. The evaluation is divided into three parts with increasing realism: in Section V-D three sources activate successively at random times with a minimum time gap of 500 ms, and oracle knowledge of source activity is assumed. In contrast, in Section V-E the DNN-based online source counting method proposed in [45] is used to estimate the activations of new sources, for the same acoustic scenario with

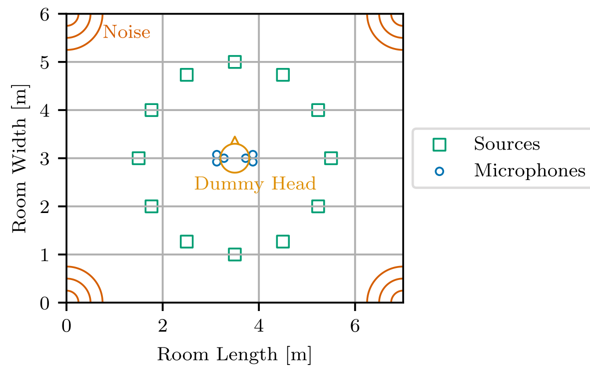

Fig. 2. Acoustic setup with 12 speaker positions, 2 behind-the-ear hearing aids with two microphones each worn by a dummy head, one in-ear microphone on each side of the dummy head, and background noise from the BRUDEX database [47].

three successively activating sources. Finally, in Section V-F the performance is evaluated with and without oracle knowledge of source activity in a more realistic scenario with three sources activating and deactivating at random times, with a minimum time gap of 50 ms between activation or deactivation events. The acoustic scenario for the first two experiments involving only successive source activations is introduced in Section V-A, whereas the specific dynamic scenario involving source deactivations is detailed later in Section V-F.

A. Acoustic Scenario

To generate noisy and reverberant microphone signals, we used measured room impulse responses (RIRs) and noise signals from the BRUDEX database [47] at a sampling frequency of 16 kHz. The acoustic setup is illustrated in Fig. 2. A dummy head was positioned approximately in the center of an acoustic laboratory measuring $7 \text{ m} \times 6 \text{ m} \times 2.7 \text{ m}$, with a reverberation time $T_{60} \approx 310$ ms. The dummy head was equipped with left and right behind-the-ear hearing aids, each featuring two microphones, and one in-ear microphone on each side, resulting in $M = 6$ microphones. Each acoustic scenario involved $K_{\text{max}} = 3$ speakers of random sex, activating successively with a random time gap between activations uniformly distributed within $[0.5 \text{ s}, 5 \text{ s}]$ and continuously active background noise (similarly as in Fig. 1). This resulted in noise-only, single-source, dual-source, and triple-source segments with random lengths. The source components at the microphones were generated by convolving clean speech signals from the Librispeech ASR Corpus [48] with RIRs measured from loudspeakers placed at 12 different positions around the dummy head in a circle with a radius of 2 m and an angular spacing of 30° (see Fig. 2). Quasi-diffuse noise was generated by playing back uncorrelated babble noise or cafeteria noise using four loudspeakers facing the corners of the laboratory. Using only the active samples for each source, the sources were normalized to have equal power, averaged across all microphones. The noise was then scaled to set the SNR, defined as the power ratio between a single source component and the noise component, to a value randomly selected from $\{0, 5, 10, 15\}\, dB$. We considered 2640 different

TABLE I
OVERVIEW OF CONSIDERED BOP-BASED RTF VECTOR ESTIMATION METHODS, RESULTING FROM COMBINATIONS OF NOISE HANDLING AND ADDITIONAL VECTORS. HERE BOP REFERS TO THE CONVENTIONAL METHOD [37] AND $*$ DENOTES THE PROPOSED METHODS

BOP-based Methods			
Additional Vectors	**Noise Handling**		
	no	subtraction	whitening
random	BOP [37]	BOP-S*	BOP-W*
orthogonal	BOPO*	BOPO-S*	BOPO-W*

scenarios with random source positions, clean speech signals, noise types, SNRs, and random source activation times.

B. Algorithmic Implementation

The algorithms were implemented using an STFT framework with a frame length of 1024 samples (corresponding to 64 ms), a frame shift of 256 samples (corresponding to $t_{\text{fs}} = 16$ ms) and a square-root-Hann window for analysis and synthesis, leading to an algorithmic latency of 64 ms. The noise covariance matrix is estimated as the sample covariance matrix in the noise-only segment $\mathcal{T}_n = [0, t_1[$, i.e., $\widehat{\mathbf{R}}_n = \frac{1}{|\mathcal{T}_n|} \sum_{t \in \mathcal{T}_n} \mathbf{y}_t \mathbf{y}_t^H$. The noisy covariance matrix $\widehat{\mathbf{R}}_{y,t}$ is initialized in the first time frame and then updated recursively using (9) with a smoothing time constant of $t_\alpha = 0.5$ s. The RTF vectors of the first, second and third source $\widehat{\mathbf{g}}_{1,t}$, $\widehat{\mathbf{g}}_{2,t}$ and $\widehat{\mathbf{g}}_{3,t}$ are estimated in the single-source, dual-source and triple-source segments, respectively. Table I provides an overview of the six BOP-based RTF vector estimation methods, comprising the conventional method [37] and the five proposed extensions formed by combining different additional vectors (random vs. orthogonal) and noise handling techniques (none, subtraction, or whitening). The matrix $\mathbf{G}_{\bar{K}}$ used by BOP-based methods is constructed using the estimated RTF vectors of the already active sources from the last frame of the corresponding segment, i.e., $\mathbf{G}_1 \in \mathbb{C}^{M \times 0}$ (empty matrix), $\mathbf{G}_2 = \widehat{\mathbf{g}}_{1,(\widehat{t_2}-1)}$ and $\mathbf{G}_3 = [\widehat{\mathbf{g}}_{1,(\widehat{t_2}-1)} \;\; \widehat{\mathbf{g}}_{2,(\widehat{t_3}-1)}]$. Note that to estimate the RTF vectors $\widehat{\mathbf{g}}_{2,t}$ and $\widehat{\mathbf{g}}_{3,t}$, each method utilizes its own previously obtained estimates of the already active sources, i.e., $\widehat{\mathbf{g}}_{1,(\widehat{t_2}-1)}$ and $\widehat{\mathbf{g}}_{2,(\widehat{t_3}-1)}$.

For all BOP-based RTF vector estimation methods (including conventional BOP), we utilized the proposed closed-form solution in (30), significantly reducing computation time compared to the iterative gradient descent optimization scheme used in [37]. In addition, the closed-form solution guarantees computing the global minimum, whereas gradient descent may converge to a local minimum and depends on multiple parameters, such as initialization, learning rate, and re-initialization scheme. Moreover, the closed-form solution is easily parallelizable across frequencies, unlike gradient descent. On an NVIDIA RTX A6000 graphics card, the proposed closed-form solution achieves an overall computation time speed-up of approximately 1×10^6.

In addition to the conventional BOP method, we consider two additional baseline methods: the CWu method [20] and the GSS method [32]. The CWu method estimates the RTF vector of the target source by utilizing an estimate of the undesired covariance matrix $\widehat{\mathbf{R}}_v$ from the previous segment, capturing both the background noise and the already active sources. Similar to the proposed whitening-based methods, the CWu method employs a whitening operation, principal eigenvector extraction, and a de-whitening step. However, instead of blocking the already active sources, the sum of all already active sources and background noise, represented by $\widehat{\mathbf{R}}_v$, is used for whitening. The RTF vector of the K-th source is estimated as

$$\widehat{\mathbf{g}}_K = \frac{\widehat{\mathbf{R}}_v^{\frac{\mathrm{H}}{2}} \tilde{\mathbf{g}}_K}{\mathbf{e}_r^{\mathrm{T}} \widehat{\mathbf{R}}_v^{\frac{\mathrm{H}}{2}} \tilde{\mathbf{g}}_K}, \quad \text{with} \quad \tilde{\mathbf{g}}_K = \mathcal{V}_{\max}\left\{ \widehat{\mathbf{R}}_v^{\frac{-\mathrm{H}}{2}} \widehat{\mathbf{R}}_y \widehat{\mathbf{R}}_v^{\frac{-1}{2}} \right\}. \tag{36}$$

The performance of all BOP-based methods and the CWu method will be evaluated in terms of the RTF vector estimation accuracy and the performance of an LCMV beamformer aiming at extracting a target source based on the estimated RTF vectors (see Section V-C). As a baseline method directly extracting a target source we consider the GSS method [32], which uses diarization labels to estimate covariance matrices of the individual sources using a complex angular central Gaussian mixture model as a generative model. The covariance matrices are then used in a Souden MVDR beamformer [33] to extract the target source. To strictly meet the latency requirement of 64 ms, we employed the block-online GSS variant from [34], hereafter referred to as GSS, with specific adjustments of its parameters: First, to minimize delay, we reduced the block length to 4 frames (64 ms). Second, to ensure consistent latency across all frames, we applied the beamformer computed in the current block only to its last frame, processing the preceding 3 frames with the beamformer from the previous block. In addition, consistent with [34], we utilized a pre-context of 50 frames (800 ms). Although the GSS method does not explicitly estimate RTF vectors, they can be extracted as the principal eigenvectors of the respective source covariance matrices. Therefore, we also assess the performance of the GSS method in terms of RTF vector estimation.

C. Performance Metrics

The RTF vector estimation accuracy is evaluated using the Hermitian angle between the estimated RTF vectors and the ground-truth RTF vectors. The Hermitian angle measures the directional similarity between two vectors and is invariant under arbitrary complex scaling of either vector. A weighted sum over all F frequency bins and all time frames $t \in \mathcal{T}_K$ is computed ($K \in \{1, 2, 3\}$), i.e.,

$$\psi_K = \frac{\sum_{t \in \mathcal{T}_K} \sum_{f=1}^{F} w_{K,f} \arccos\left\{ \frac{|\mathbf{g}_{K,f}^{\mathrm{H}} \widehat{\mathbf{g}}_{K,t,f}|}{\|\mathbf{g}_{K,f}\| \|\widehat{\mathbf{g}}_{K,t,f}\|} \right\}}{|\mathcal{T}_K| \sum_{f=1}^{F} w_{K,f}}, \tag{37}$$

where the vectors $\mathbf{g}_{K,f}$ and $\widehat{\mathbf{g}}_{K,t,f}$ denote the ground-truth and the estimated RTF vector in frequency bin f, respectively. The ground-truth RTF vector of the K-th source is computed as the principal eigenvector of its sample covariance matrix, i.e., $\mathbf{g}_{K,f} = \mathcal{V}_{\max}\{\mathbf{R}_{K,f}\}$. The frequency weights are set to the average signal power of the K-th source, i.e., $w_{K,f} = \mathrm{tr}\{\mathbf{R}_{K,f}\}$ with $\mathbf{R}_{K,f} = \frac{1}{|\mathcal{T}_K|} \sum_{t \in \mathcal{T}_K} \mathbf{x}_{K,t,f} \mathbf{x}_{K,t,f}^{\mathrm{H}}$. This weighting ensures that Hermitian angles in frequency bins with low signal power, which are less relevant for signal enhancement, are weighted less. Note that lower values of the weighted Hermitian angle in (37) indicate better performance.

To evaluate the practical benefit of the estimated RTF vectors for speech enhancement as a potential application, we also consider the broadband signal-to-interferer-and-noise ratio (SINR) improvement of an LCMV beamformer [2], [49], [50]. The LCMV beamformer utilizes the estimated RTF vectors to extract a target source while suppressing interfering sources and background noise. In our evaluation, we consistently define the target source as the most recently activated source among the currently active sources, i.e. the K-th source within the segment $\mathcal{T}_K$, while all other active sources are treated as interfering sources. This implies that while the target source remains fixed within a segment, it changes dynamically over time as sources activate (or deactivate in the experiment described in Section V-F). The LCMV beamformer minimizes the noise power spectral densities (PSDs) subject to linear constraints, which aim at extracting the K-th source component in the reference microphone without distortion and suppressing the $K-1$ interfering components by a pre-defined amount, i.e.,

$$\mathbf{w}_K = \arg\min_{\mathbf{w}} \left(\mathbf{w}^{\mathrm{H}} \mathbf{R}_n \mathbf{w} \right) \quad \text{s.t.} \quad \begin{array}{c} \mathbf{w}^{\mathrm{H}} \mathbf{x}_{K,t} = x_{K,r,t} \\ \mathbf{w}^{\mathrm{H}} \mathbf{x}_{k,t} = \delta x_{k,r,t} \\ \forall k \in \{1, \ldots, K-1\} \end{array}, \tag{38}$$

where $\delta \in [0, 1[$ denotes the interference suppression factor (assumed to be equal for all interfering sources). By reformulating the linear constraints in terms of the RTF vectors, the well-known solution to (38) is given by

$$\boxed{\mathbf{w}_K = \mathbf{R}_n^{-1} \mathbf{C}_K \left(\mathbf{C}_K^{\mathrm{H}} \mathbf{R}_n^{-1} \mathbf{C}_K \right)^{-1} \boldsymbol{\delta}} \tag{39}$$

where the constraint matrix $\mathbf{C}_K \in \mathbb{C}^{M \times K}$ contains the RTF vectors of all K active sources, i.e., $\mathbf{C}_K = \mathbf{g}_K \quad \mathbf{G}_{\bar{K}}$, and $\boldsymbol{\delta} = [1 \quad \delta \quad \cdots \quad \delta]^{\mathrm{T}} \in \mathbb{R}_+^{K \times 1}$ is the interference suppression vector. Applying the LCMV beamformer to the microphone signals in the K-th segment yields the output signal

$$\mathbf{z}_t = \mathbf{w}_K^{\mathrm{H}} \mathbf{y}_t \quad \text{for} \quad t_K \leq t < t_{K+1}. \tag{40}$$

The LCMV beamformer in (39) with $\delta = -20$ dB is applied to the K-th source component $x_{r,K,t}$ and the K-th undesired component $v_{r,\bar{K},t}$ in the STFT domain, selecting the front microphone on the left hearing aid as the reference microphone ($r = 1$). The corresponding single-channel time-domain output signals $x_{r,K,l}^{\mathrm{out}} \in \mathbb{R}$ and $v_{r,\bar{K},l}^{\mathrm{out}} \in \mathbb{R}$, with l denoting the time-domain sample index, are computed by applying the inverse STFT in an overlap-add procedure. The single-channel broadband SINR improvement in the K-th segment is defined as

$$\Delta\mathrm{SINR}_K = 10\log_{10}\left\{ \frac{\sum_{l \in \mathcal{L}_K} |x_{r,K,l}^{\mathrm{out}}|^2}{\sum_{l \in \mathcal{L}_K} |v_{r,\bar{K},l}^{\mathrm{out}}|^2} \right\} - 10\log_{10}\left\{ \frac{\sum_{l \in \mathcal{L}_K} |x_{r,K,l}|^2}{\sum_{l \in \mathcal{L}_K} |v_{r,\bar{K},l}|^2} \right\}, \tag{41}$$

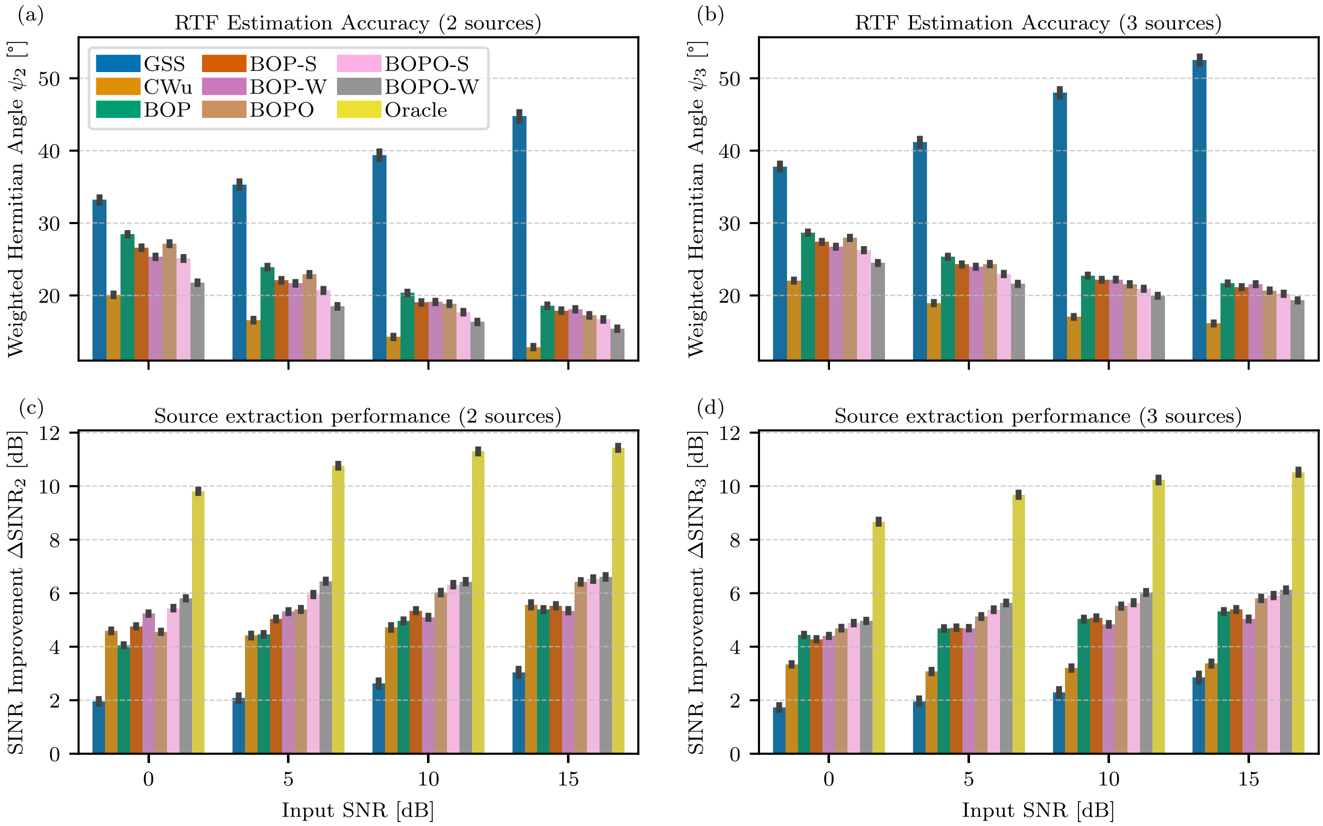

Fig. 3. Performance evaluation of the conventional and proposed RTF vector estimation methods for different SNRs using oracle source activity knowledge in terms of RTF vector estimation accuracy and SINR improvement. (a) and (c) show the performance for the dual-source segments, while (b) and (d) show the performance for the triple-source segments.

TABLE II
MEAN REAL-TIME FACTOR FOR ALL CONSIDERED METHODS ON A SINGLE NVIDIA RTX A5000 GPU

Method	**Mean Real-time Factor**
BOP [37]	0.0044
BOP-S	0.0045
BOP-W	0.0048
BOPO	0.0077
BOPO-S	0.0079
BOPO-W	0.0082
CWu [20]	0.0072
GSS [34]	0.1431

where $x_{r,K,l}$ and $v_{r,\bar{K},l}$ denote the time-domain K-th source component and undesired components in the reference microphone, and $\mathcal{L}_K$ denotes the set of samples within the K-th segment.

D. Results Using Oracle Source Activity Knowledge

In this section, we investigate the benefit of the extensions proposed in Sections IV-B and IV-C compared to the conventional BOP method and other baseline methods in scenarios with only source activations, assuming oracle source activity knowledge.

Before analyzing the performance, we evaluate the computational complexity. For all considered methods, Table II presents the mean real-time factor measured on a single NVIDIA RTX A5000 graphics card. All BOP-based methods utilizing the proposed closed-form solution demonstrate high efficiency with real-time factors below 0.005. While the use of orthogonal additional vectors in BOPO leads to a minor increase in complexity (≈ 0.008), it remains comparable to the baseline CWu method. In contrast, the GSS method exhibits a significantly higher computational complexity, approximately 20 times higher than the BOPO methods.

For different SNRs, Fig. 3 depicts the weighted Hermitian angle for all considered RTF vector estimation methods in dual-source and triple-source segments. In addition, the figure shows the SINR improvement of the LCMV beamformer using the estimated RTF vectors and the baseline GSS method, in dual-source and triple-source segments. The performance of the LCMV beamformer using oracle RTF vectors is included for comparison. In addition, the median values are provided in Table III. The barplots show the median over all evaluated scenarios with confidence interval ($p = 0.05$). First, it can be observed that in general the performance of all considered RTF vector estimation methods improves (i.e., smaller weighted Hermitian angle and larger SINR improvement) with increasing SNR, both for the dual-source and triple-source segments. In addition, for all SNRs all considered methods perform better in the dual-source segments compared to the triple-source segments. Second, we compare the performance between using orthogonal additional vectors (BOPO methods) and random

TABLE III
MEDIAN WEIGHTED HERMITIAN ANGLE ψ AND MEDIAN SINR IMPROVEMENT FOR THE CONVENTIONAL AND THE PROPOSED METHODS FOR DIFFERENT SNRS ASSUMING ORACLE SOURCE ACTIVITY KNOWLEDGE

Method	SNR 0		SNR 5		SNR 10		SNR 15	
	$\psi\downarrow$	ΔSINR↑	$\psi\downarrow$	ΔSINR↑	$\psi\downarrow$	ΔSINR↑	$\psi\downarrow$	ΔSINR↑
2 sources								
GSS [34]	33.2°	2.0 dB	35.3°	2.1 dB	39.4°	2.6 dB	44.7°	3.0 dB
CWu [20]	**20.1°**	4.6 dB	**16.6°**	4.4 dB	**14.3°**	4.7 dB	**12.9°**	5.6 dB
BOP [37]	28.4°	4.1 dB	23.9°	4.5 dB	20.3°	5.0 dB	18.6°	5.4 dB
BOP-S*	26.6°	4.8 dB	22.1°	5.0 dB	19.0°	5.3 dB	17.9°	5.5 dB
BOP-W*	25.3°	5.2 dB	21.7°	5.3 dB	19.1°	5.1 dB	18.1°	5.3 dB
BOPO*	27.1°	4.6 dB	22.9°	5.4 dB	18.9°	6.0 dB	17.2°	6.4 dB
BOPO-S*	25.1°	5.4 dB	20.7°	6.0 dB	17.7°	6.3 dB	16.7°	6.5 dB
BOPO-W*	21.8°	**5.8 dB**	18.5°	**6.4 dB**	16.3°	**6.4 dB**	15.4°	**6.6 dB**
Oracle	0.0°	9.8 dB	0.0°	10.8 dB	0.0°	11.3 dB	0.0°	11.4 dB
3 sources								
GSS [34]	37.8°	1.7 dB	41.2°	2.0 dB	48.0°	2.3 dB	52.5°	2.8 dB
CWu [20]	**22.0°**	3.3 dB	**18.9°**	3.1 dB	**17.0°**	3.2 dB	**16.1°**	3.4 dB
BOP [37]	28.7°	4.4 dB	25.3°	4.7 dB	22.7°	5.0 dB	21.7°	5.3 dB
BOP-S*	27.4°	4.3 dB	24.3°	4.7 dB	22.2°	5.1 dB	21.1°	5.4 dB
BOP-W*	26.7°	4.4 dB	24.0°	4.7 dB	22.2°	4.8 dB	21.6°	5.0 dB
BOPO*	27.9°	4.7 dB	24.3°	5.1 dB	21.5°	5.5 dB	20.7°	5.8 dB
BOPO-S*	26.3°	4.9 dB	23.0°	5.4 dB	20.9°	5.6 dB	20.2°	5.9 dB
BOPO-W*	24.5°	**5.0 dB**	21.6°	**5.6 dB**	20.0°	**6.0 dB**	19.3°	**6.1 dB**
Oracle	0.0°	8.7 dB	0.0°	9.7 dB	0.0°	10.2 dB	0.0°	10.5 dB

additional vectors (BOP methods). For all SNRs and noise handling techniques, the results for the dual-source segments in Fig. 3(a) and the triple-source segments in Fig. 3(b) show that orthogonal additional vectors outperform random additional vectors in terms of weighted Hermitian angle, especially at low SNRs. The advantage of orthogonal over random additional vectors is also observed in terms of SINR improvement, see Fig. 3(c) and 3(d). Third, we compare the performance between different noise handling techniques. Incorporating noise handling for the baseline BOP method with random additional vectors significantly improves performance (i.e. smaller weighted Hermitian angle and larger SINR improvement), where noise whitening outperforms noise subtraction for low SNR conditions and noise subtraction outperforms noise whitening for high SNR conditions. Also when using orthogonal additional vectors, noise handling significantly improves performance, especially for low SNR conditions, where noise whitening outperforms noise subtraction. Comparing the BOP-based methods, the results in Fig. 3 show that the BOPO-W method combining orthogonal additional vectors and noise whitening outperforms the other BOP-based methods in terms of SINR improvement and RTF vector estimation accuracy. For example, for the challenging 0 dB SNR condition in the dual-source segments, the proposed BOPO-W method yields a median improvement of 6.6° in terms of weighted Hermitian angle and 1.7 dB in terms of SINR improvement compared to the conventional BOP method.

Finally, we compare the performance of the proposed methods to the baseline GSS and CWu methods and the upper benchmark using oracle RTF vectors in the LCMV beamformer. The block-online GSS method, yields the largest weighted Hermitian angle and the lowest SINR improvement of all considered methods. This is likely due to the short block length of 64 ms required to meet the low latency requirement. The CWu method yields the lowest weighted Hermitian angle of all RTF vector estimation methods. However, in terms of SINR improvement, the CWu method performs significantly worse than the proposed BOPO methods (e.g., 4.6 dB vs 5.8 dB at 0 dB SNR in the dual-source segments). This discrepancy indicates that a lower Hermitian angle does not always translate to a higher SINR improvement. In terms of SINR improvement, it can be observed that there is still a performance gap between the proposed BOPO-W method and the LCMV beamformer using oracle RTF vectors across all SNR conditions (e.g., 4 dB at 0 dB SNR in the dual-source segments). This indicates that there is still potential for further improvement in RTF vector estimation methods.

In conclusion, the results in Fig. 3 show that the proposed BOPO-W method combining orthogonal additional vectors and noise whitening provides the best performance of all considered methods in terms of SINR improvement while also achieving competitive RTF vector estimation accuracy.

E. Results Using Online Source Counting

Instead of assuming oracle source activity knowledge, in this section we compare the performance of all considered RTF vector estimation methods using the DNN-based online source counting method proposed in [45]. This method estimates the number of active sources on a frame-by-frame basis using the generalized magnitude-squared coherence (GMSC) of a whitened version of the noisy covariance matrix as input features for a DNN classifier. Although the method provides estimates of the number of active sources every 16 ms (corresponding to the frame shift), the reliance on recursively smoothed covariance matrices with a 500 ms time constant inherently introduces an implicit delay before a newly active source is sufficiently

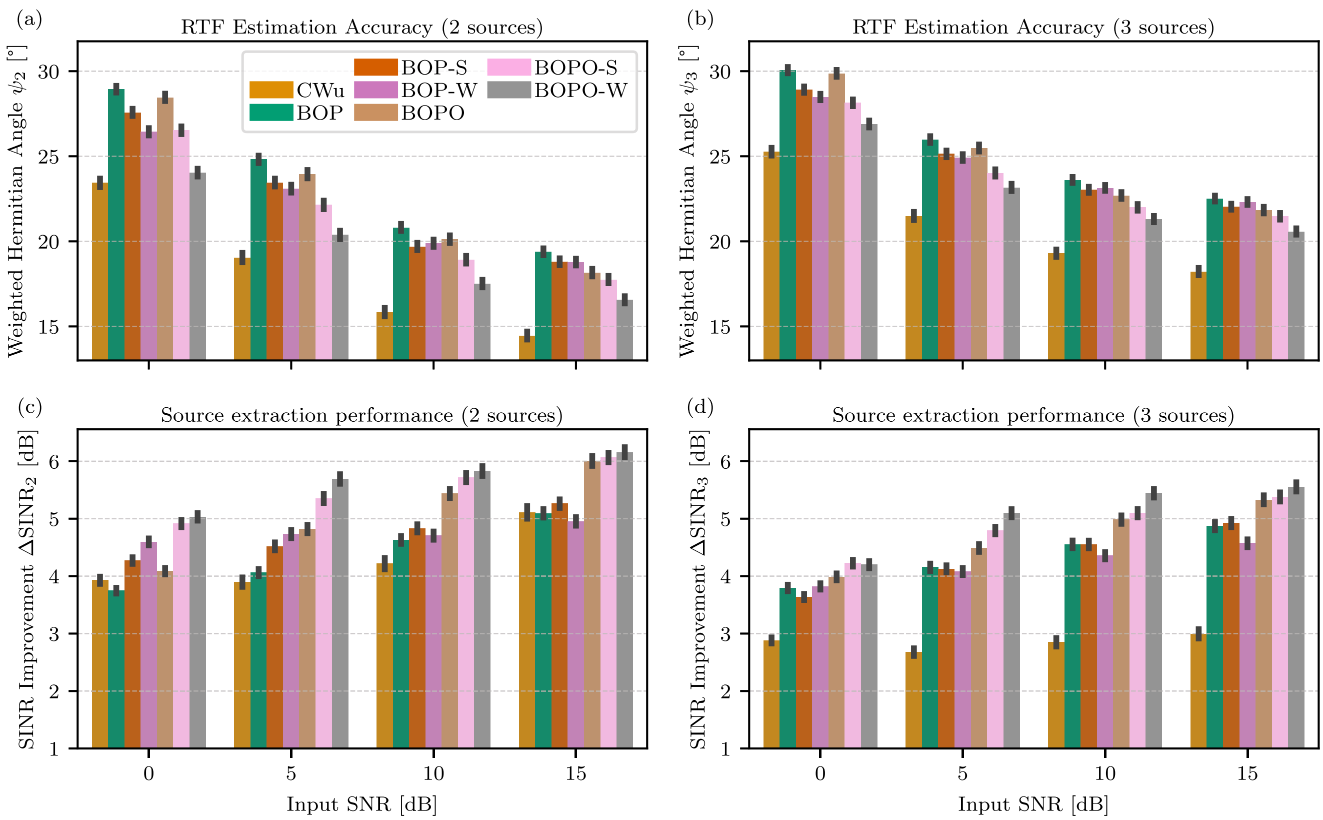

Fig. 4. Performance evaluation of the conventional and proposed RTF vector estimation methods for different SNRs using the online source counting method from [45] in terms of RTF vector estimation accuracy and SINR improvement. (a) and (c) show the performance for the dual-source segments, while (b) and (d) show the performance for the triple-source segments.

reflected in the input features. Here we used the recurrent neural network (RNN)-based architecture proposed in [45], which consists of 3 gated recurrent unit (GRU) layers with a hidden size of half the input dimension. The DNN classifier is trained on 5280 training scenarios and 528 validation scenarios. This online source counting method achieved a frame-level accuracy of 95% on the test scenarios. It should be noted that the GSS method is not considered in this section, as it depends on diarization labels, which are not provided by the online source counting method.

For different SNRs, Fig. 4 depicts the weighted Hermitian angle and the SINR improvement in dual-source and triple-source segments (median values are provided in Table IV). First, it can be observed that the performance for all considered methods and conditions is only slightly lower when using the GMSC-based online source counting method compared to assuming oracle source activity knowledge. For the weighted Hermitian angle, the maximal degradation is 3.2° and for the SINR improvement, the maximal degradation is 0.8 dB. The overall performance trends when using the online source counting method are very similar to those when assuming oracle source activity knowledge: orthogonal additional vectors outperform random additional vectors, and noise whitening outperforms noise subtraction and no noise handling, with larger benefits for lower SNRs. The baseline CWu method shows similar performance as when using oracle source activity knowledge, again yielding the lowest weighted Hermitian angle, but performing significantly worse than the proposed BOPO methods in terms of SINR improvement. In conclusion, the proposed BOPO-W method combining orthogonal additional vectors and noise whitening also provides the best performance of all considered methods in terms of SINR improvement while also achieving competitive RTF vector estimation accuracy when using online source counting.

F. Dynamic Scenario With Source Deactivations

To evaluate the performance of all considered methods in more realistic scenarios, in this section we consider scenarios where up to three sources can activate and deactivate within an utterance. We use the same acoustic setup and dataset as described in Section V-A, but now we construct scenarios with a fixed signal length of 60 s. The time intervals between two consecutive activation or deactivation events are randomly chosen between 50 ms and 5 s, allowing for much shorter segment lengths compared to the previous experiment (Sections V-D and V-E). At each activation event, a new source activates (up to a maximum of three active sources), while at each deactivation event, one of the currently active sources is randomly chosen to deactivate. Thus, at any time instant, there can be zero, one, two, or three active sources. For this experiment, we randomly constructed 750 test scenarios. We provide results assuming oracle source activity knowledge, as

TABLE IV
MEDIAN WEIGHTED HERMITIAN ANGLE ψ AND SINR IMPROVEMENT FOR THE CONVENTIONAL AND PROPOSED RTF VECTOR ESTIMATION METHODS FOR DIFFERENT SNRS USING THE ONLINE SOURCE COUNTING METHOD [45]

Method	SNR 0		5		10		15	
	$\psi\downarrow$	ΔSINR↑	$\psi\downarrow$	ΔSINR↑	$\psi\downarrow$	ΔSINR↑	$\psi\downarrow$	ΔSINR↑
2 sources								
CWu [20]	**23.4°**	3.9 dB	**19.0°**	3.9 dB	**15.8°**	4.2 dB	**14.5°**	5.1 dB
BOP [37]	28.9°	3.7 dB	24.8°	4.1 dB	20.8°	4.6 dB	19.4°	5.1 dB
BOP-S*	27.6°	4.3 dB	23.5°	4.5 dB	19.7°	4.8 dB	18.8°	5.3 dB
BOP-W*	26.4°	4.6 dB	23.1°	4.7 dB	19.9°	4.7 dB	18.8°	5.0 dB
BOPO*	28.5°	4.1 dB	23.9°	4.8 dB	20.1°	5.4 dB	18.2°	6.0 dB
BOPO-S*	26.5°	4.9 dB	22.1°	5.4 dB	18.9°	5.7 dB	17.8°	6.1 dB
BOPO-W*	24.0°	**5.0 dB**	20.4°	**5.7 dB**	17.5°	**5.8 dB**	16.6°	**6.2 dB**
3 sources								
CWu [20]	**25.3°**	2.9 dB	**21.5°**	2.7 dB	**19.3°**	2.8 dB	**18.2°**	3.0 dB
BOP [37]	30.1°	3.8 dB	26.0°	4.2 dB	23.6°	4.6 dB	22.5°	4.9 dB
BOP-S*	28.9°	3.6 dB	25.1°	4.1 dB	23.0°	4.6 dB	22.0°	4.9 dB
BOP-W*	28.5°	3.8 dB	24.9°	4.1 dB	23.1°	4.4 dB	22.3°	4.6 dB
BOPO*	29.9°	4.0 dB	25.5°	4.5 dB	22.7°	5.0 dB	21.8°	5.3 dB
BOPO-S*	28.1°	**4.2 dB**	24.0°	4.8 dB	22.0°	5.1 dB	21.5°	5.4 dB
BOPO-W*	26.9°	**4.2 dB**	23.2°	**5.1 dB**	21.3°	**5.4 dB**	20.6°	**5.6 dB**

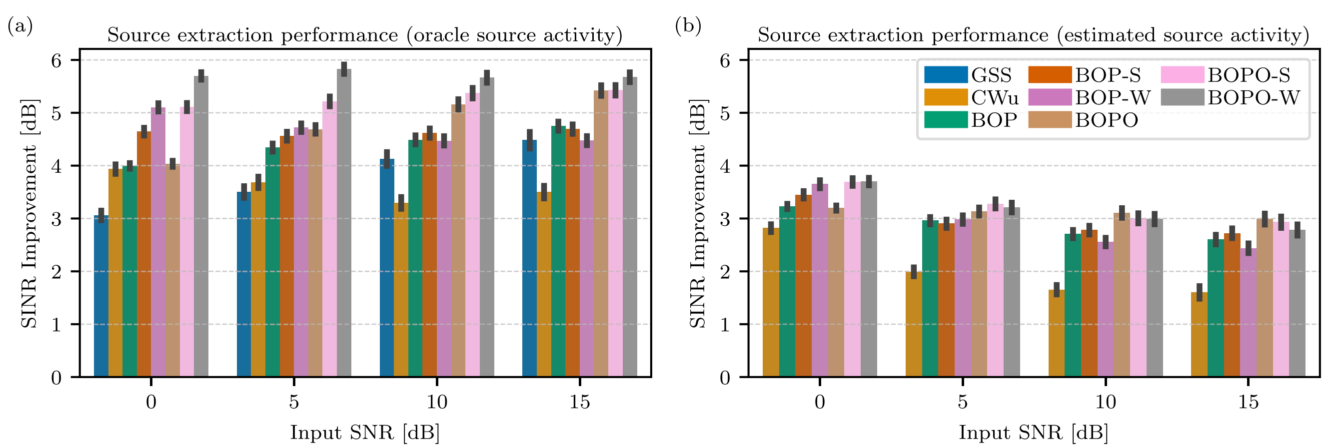

Fig. 5. Performance evaluation of all considered methods in terms of SINR improvement for different SNRs in dynamic scenarios with source activations and deactivations. (a) using oracle source activity knowledge, (b) using the DNN-based online source counting method in [45].

well as results using the online source counting method from [45] (using the same configuration as in Section V-E and trained on 1500 training and 150 validation scenarios). In this experiment with source deactivations, the online source counting method achieved a frame-level accuracy of 87.4% on the test scenarios. In segments where a source deactivates (i.e., K decreases), no new RTF vector needs to be estimated. Instead, the matrix $\mathbf{G}_{\bar{K}}$ is updated by removing the RTF vector of the deactivated source. To identify this source online, we estimate the output power of K internal linearly constrained minimum power (LCMP) beamformers extracting each of the currently active sources. The RTF vector corresponding to the lowest output power is then removed. As in previous experiments, the target source is defined as the most recently activated source among those remaining active. The performance is evaluated in terms of the utterance-level SINR improvement, averaged over all segments.

For different SNRs, Fig. 5 depicts the SINR improvement of all considered methods for the activation and deactivation scenario, either using oracle or estimated source activity. The median values are provided in Table V. Similarly as in Section V-E, the GSS method is not considered when using estimated source activity. Assuming oracle source activity knowledge, the results in Fig. 5(a) exhibit very similar trends to the main experiment described in Section V-D. The proposed BOPO methods generally outperform the baseline RTF vector estimation methods (CWu, BOP) and the GSS method, with the BOPO-W method combining orthogonal additional vectors and noise whitening achieving the highest SINR improvement across all SNRs (e.g., 5.7 dB in the challenging 0 dB SNR condition). Also consistent with previous findings, noise handling proves more effective at lower SNRs with noise whitening outperforming noise subtraction, while orthogonal additional vectors

TABLE V
MEDIAN SINR IMPROVEMENT OF ALL CONSIDERED METHODS FOR DIFFERENT SNRS USING ORACLE AND ESTIMATED SOURCE ACTIVITY

Method	SNR 0	5	10	15
Oracle Source Activity				
GSS [34]	3.1 dB	3.5 dB	4.1 dB	4.5 dB
CWu [20]	3.9 dB	3.7 dB	3.3 dB	3.5 dB
BOP [37]	4.0 dB	4.3 dB	4.5 dB	4.7 dB
BOP-S*	4.6 dB	4.6 dB	4.6 dB	4.7 dB
BOP-W*	5.1 dB	4.7 dB	4.5 dB	4.5 dB
BOPO*	4.0 dB	4.7 dB	5.2 dB	5.4 dB
BOPO-S*	5.1 dB	5.2 dB	5.4 dB	5.4 dB
BOPO-W*	**5.7 dB**	**5.8 dB**	**5.7 dB**	**5.7 dB**
Estimated Source Activity				
CWu [20]	2.8 dB	2.0 dB	1.7 dB	1.6 dB
BOP [37]	3.2 dB	3.0 dB	2.7 dB	2.6 dB
BOP-S*	3.4 dB	2.9 dB	2.8 dB	2.7 dB
BOP-W*	3.6 dB	3.0 dB	2.6 dB	2.4 dB
BOPO*	3.2 dB	3.1 dB	**3.1 dB**	**3.0 dB**
BOPO-S*	**3.7 dB**	**3.3 dB**	3.0 dB	2.9 dB
BOPO-W*	**3.7 dB**	3.2 dB	3.0 dB	2.8 dB

provide a robust advantage over random additional vectors. When using the online source counting method, the results in Fig. 5(b) show a notable performance degradation and less distinct trends compared to the oracle case. This degradation can be attributed to the increased difficulty of estimating the source activity in this dynamic scenario, given the incorporation of deactivations and the short minimum segment length of 50 ms compared to 500 ms in the previous experiment. Consequently, the benefits of noise handling become less pronounced, especially at higher SNRs where the BOPO method without noise handling appears most robust. Nevertheless, the proposed BOPO-based methods still consistently outperform the conventional CWu and BOP methods.

VI. CONCLUSION

In this paper, we proposed three extensions to the conventional BOP method to improve successive multi-source RTF vector estimation. First, we derived a closed-form solution by setting the vector gradient of the BOP cost function equal to zero, significantly reducing the computational complexity of the BOP method compared to using an iterative gradient descent method. Second, we provided a deeper understanding of the need and impact of additional vectors in the optimization process, and proposed to use orthogonal additional vectors instead of random additional vectors in order to reduce undesired amplification of error signal components caused by model mismatch. Third, we incorporated two noise handling techniques, namely noise subtraction and noise whitening, assuming an estimate of the noise covariance matrix is available. Both when assuming oracle source activity knowledge as well as when using an online source counting method, simulation results in noisy reverberant scenarios with sources activating and deactivating with minimum time gaps of 50 ms between two consecutive activation or deactivation events demonstrate that the proposed BOPO-W method, which combines the closed-form solution, orthogonal additional vectors, and noise whitening, results in substantial improvements in terms of computational complexity, RTF vector estimation accuracy and SINR improvement compared to the conventional BOP method. Furthermore, the proposed method yields the highest SINR improvement among all considered baseline methods, significantly outperforming the CWu method despite its slightly better RTF vector estimation accuracy. These results demonstrate the robustness and practicality of our proposed extensions for multi-source RTF vector estimation in scenarios with dynamic source activity patterns and low SNRs. Future work will focus on eliminating the need for additional vectors, and adapting the proposed methods to scenarios with moving sources.

APPENDIX A
DERIVATION OF VECTOR GRADIENT

The vector gradient $\nabla_{\boldsymbol{\theta}} J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta})$ of the BOP cost function in (17) can be derived using complex-valued matrix differentiation [51]. The gradient of a scalar real-valued function is given by $\nabla_{\boldsymbol{\theta}} J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}) = (\mathcal{D}_{\boldsymbol{\theta}} J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}))^{\mathrm{H}}$, where $\mathcal{D}_{\boldsymbol{\theta}}$ denotes the complex-valued matrix derivative operator with respect to the vector variable $\boldsymbol{\theta}$. Inserting the BOP cost function in (17), applying the chain rule described in [51, Sec. 3.4.1] and using [51, Example 4.13] yields

$$\begin{aligned}\mathcal{D}_{\boldsymbol{\theta}} J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}) &= \mathcal{D}_{\boldsymbol{\theta}} \mathrm{tr}\{\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} \mathbf{R}_{x,t} \mathbf{P}^{\angle \mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\} \\ &= \mathrm{vec}^{\mathrm{T}}\{\mathbf{I}_M\}\, \mathcal{D}_{\boldsymbol{\theta}} \left(\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} \mathbf{R}_{x,t} \mathbf{P}^{\angle \mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right), \end{aligned} \tag{42}$$

where $\mathrm{vec}\{\cdot\}$ denotes the vectorization operator, which stacks the columns of a matrix into a column vector. Applying the product rule from [51, Lemma 3.4] to (42) leads to

$$\begin{aligned}\mathcal{D}_{\boldsymbol{\theta}}\left(\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} \mathbf{R}_{x,t} \mathbf{P}^{\angle \mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right) &= \left(\left(\mathbf{R}_{x,t} \mathbf{P}^{\angle \mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right)^{\mathrm{T}} \otimes \mathbf{I}_M\right) \mathcal{D}_{\boldsymbol{\theta}} \mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} \\ &+ \left(\mathbf{I}_M \otimes \left(\mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} \mathbf{R}_{x,t}\right)\right) \mathcal{D}_{\boldsymbol{\theta}} \mathbf{P}^{\angle \mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}, \end{aligned} \tag{43}$$

where $\otimes$ denotes the Kronecker product. Using the chain rule, identities involving complex conjugation [51, Lemma 3.3] and results from [51, Table 4.4], the derivative of the Hermitian oblique projection operator $\mathcal{D}_{\boldsymbol{\theta}} \mathbf{P}^{\angle \mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}$ in (43) is given by

$$\mathcal{D}_{\boldsymbol{\theta}} \mathbf{P}^{\angle \mathrm{H}}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} = \mathbf{C}_{M,M} \left(\mathcal{D}_{\boldsymbol{\theta}^*} \mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}\right)^*, \tag{44}$$

where $\mathbf{C}_{M,N} \in \{0,1\}^{MN \times MN}$ denotes the commutation matrix of an $M \times N$ dimensional matrix $\mathbf{A}$, i.e., $\mathbf{C}_{M,N}\mathrm{vec}\{\mathbf{A}\} = \mathrm{vec}\{\mathbf{A}^{\mathrm{T}}\}$. The derivative of the oblique projection operator $\mathcal{D}_{\boldsymbol{\theta}} \mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}$ in (43) can be derived using the product rule, i.e.,

$$\begin{aligned}\mathcal{D}_{\boldsymbol{\theta}} \mathbf{P}^{\angle}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}} &= \mathcal{D}_{\boldsymbol{\theta}}\left(\mathbf{G}_{\bar{K}}\left(\mathbf{G}^{\mathrm{H}}_{\bar{K}} \mathbf{P}^{\perp}_{\boldsymbol{\theta}} \mathbf{G}_{\bar{K}}\right)^{-1} \mathbf{G}^{\mathrm{H}}_{\bar{K}} \mathbf{P}^{\perp}_{\boldsymbol{\theta}}\right) \\ &= \left(\left(\mathbf{G}^{\mathrm{H}}_{\bar{K}} \mathbf{P}^{\perp}_{\boldsymbol{\theta}}\right)^{\mathrm{T}} \otimes \mathbf{G}_{\bar{K}}\right) \mathcal{D}_{\boldsymbol{\theta}} \left(\mathbf{G}^{\mathrm{H}}_{\bar{K}} \mathbf{P}^{\perp}_{\boldsymbol{\theta}} \mathbf{G}_{\bar{K}}\right)^{-1} \\ &+ \left(\mathbf{I}_M \otimes \mathbf{G}_{\bar{K}}\left(\mathbf{G}^{\mathrm{H}}_{\bar{K}} \mathbf{P}^{\perp}_{\boldsymbol{\theta}} \mathbf{G}_{\bar{K}}\right)^{-1} \mathbf{G}^{\mathrm{H}}_{\bar{K}}\right) \mathcal{D}_{\boldsymbol{\theta}} \mathbf{P}^{\perp}_{\boldsymbol{\theta}}. \end{aligned} \tag{45}$$

Using the chain rule, the derivative of the matrix inversion [51, Example 4.23, Table 4.4] and the product rule leads to

$$\begin{aligned}&\mathcal{D}_{\boldsymbol{\theta}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\\&=-\left(\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-\mathrm{T}}\otimes\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\right)\mathcal{D}_{\boldsymbol{\theta}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)\\&=-\left(\left(\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\right)^{\mathrm{T}}\otimes\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\right)\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}. \quad (46)\end{aligned}$$

Substituting (46) into (45), using the Kronecker product identity from [51, Lemma 2.10] and merging terms yields

$$\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}=\left(\left(\mathbf{I}_M-\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}\right)^{\mathrm{T}}\otimes\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\right)\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}. \quad (47)$$

Note that no differentiation in (42)–(47) depends directly on $\boldsymbol{\theta}$, so that $\mathcal{D}_{\boldsymbol{\theta}^*}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}$ in (44) can be derived by exchanging $\mathcal{D}_{\boldsymbol{\theta}}$ with $\mathcal{D}_{\boldsymbol{\theta}^*}$ in (47), i.e.,

$$\mathcal{D}_{\boldsymbol{\theta}^*}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}=\left(\left(\mathbf{I}_M-\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}\right)^{\mathrm{T}}\otimes\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\right)\mathcal{D}_{\boldsymbol{\theta}^*}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}. \quad (48)$$

The derivative of the complement orthogonal projection operator $\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}$ in (47) can be derived by first applying the product rule, i.e.,

$$\begin{aligned}\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}&=\mathcal{D}_{\boldsymbol{\theta}}\left(\mathbf{I}_M-\boldsymbol{\theta}\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)^{-1}\boldsymbol{\theta}^{\mathrm{H}}\right)=-\left(\boldsymbol{\theta}^{+\mathrm{T}}\otimes\mathbf{I}_M\right)\mathcal{D}_{\boldsymbol{\theta}}\boldsymbol{\theta}\\&\quad-\left(\mathbf{I}_M\otimes\boldsymbol{\theta}^{+\mathrm{H}}\right)\mathcal{D}_{\boldsymbol{\theta}}\boldsymbol{\theta}^{\mathrm{H}}-\left(\boldsymbol{\theta}^*\otimes\boldsymbol{\theta}\right)\mathcal{D}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)^{-1}. \quad (49)\end{aligned}$$

Using the chain rule, the derivative of the matrix inversion in (49) is given by

$$\mathcal{D}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)^{-1}=-\left(\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)^{-1}\otimes\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)^{-1}\right)\mathcal{D}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right), \quad (50)$$

$$\text{with}\quad \mathcal{D}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\theta}\right)=\boldsymbol{\theta}^{\mathrm{T}}\mathcal{D}_{\boldsymbol{\theta}}\boldsymbol{\theta}^{\mathrm{H}}+\boldsymbol{\theta}^{\mathrm{H}}\mathcal{D}_{\boldsymbol{\theta}}\boldsymbol{\theta}, \quad (51)$$

using the product rule. Inserting (50) and (51) into (49) and evaluating $\mathcal{D}_{\boldsymbol{\theta}}\boldsymbol{\theta}^{\mathrm{H}}$ and $\mathcal{D}_{\boldsymbol{\theta}}\boldsymbol{\theta}$ according to [51, Table 4.4] yields

$$\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}=-\left(\boldsymbol{\theta}^{+\mathrm{T}}\otimes\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\right). \quad (52)$$

Using identities involving complex conjugation, $\mathbf{P}_{\boldsymbol{\theta}}^{\perp}=\mathbf{P}_{\boldsymbol{\theta}}^{\perp\mathrm{H}}=(\mathbf{P}_{\boldsymbol{\theta}}^{\perp\mathrm{T}})^*$, the chain rule, (52), and the Kronecker product in combination with commutation matrices [51, Lemma 2.12], the derivative $\mathcal{D}_{\boldsymbol{\theta}^*}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}$ in (48) can be derived as

$$\mathcal{D}_{\boldsymbol{\theta}^*}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}=\left(\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp\mathrm{T}}\right)^*=\left(\mathbf{C}_{M,M}\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\right)^*=-\left(\mathbf{P}_{\boldsymbol{\theta}}^{\perp\mathrm{T}}\otimes\boldsymbol{\theta}^{+\mathrm{H}}\right). \quad (53)$$

Inserting (52) into (47) and (53) into (48), and using the Kronecker product identity from [51, Lemma 2.10] and (10) yields

$$\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}=\left(\boldsymbol{\theta}^{+}\left(\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}-\mathbf{I}_M\right)\right)^{\mathrm{T}}\otimes\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}. \quad (54)$$

$$\mathcal{D}_{\boldsymbol{\theta}^*}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}=\left(\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\left(\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}-\mathbf{I}_M\right)\right)^{\mathrm{T}}\otimes\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\boldsymbol{\theta}^{+\mathrm{H}}. \quad (55)$$

Using $\mathbf{I}_M-\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}=\mathbf{P}_{\boldsymbol{\theta}\mathbf{G}_{\bar{K}}}^{\angle}+\mathbf{P}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}^{\perp}$ simplifies (54) and (55) to

$$\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}=-\left(\boldsymbol{\theta}^{+}\mathbf{P}_{\boldsymbol{\theta}\mathbf{G}_{\bar{K}}}^{\angle}\right)^{\mathrm{T}}\otimes\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle} \quad (56)$$

$$\mathcal{D}_{\boldsymbol{\theta}^*}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}=-\mathbf{P}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}^{\perp\mathrm{T}}\otimes\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\boldsymbol{\theta}^{+\mathrm{H}}. \quad (57)$$

Inserting (57) into (44) and using [51, Lemma 2.12] yields

$$\mathcal{D}_{\boldsymbol{\theta}}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle\mathrm{H}}=-\left(\boldsymbol{\theta}^{+}\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\right)^{\mathrm{T}}\otimes\mathbf{P}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}^{\perp}. \quad (58)$$

Inserting (43), (56), and (58) into (42) yields

$$\begin{aligned}\mathcal{D}_{\boldsymbol{\theta}}J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta})&=-\mathrm{vec}^{\mathrm{T}}\{\mathbf{I}_M\}\left(\left(\boldsymbol{\theta}^{+}\mathbf{P}_{\boldsymbol{\theta}\mathbf{G}_{\bar{K}}}^{\angle}\mathbf{R}_{x,t}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle\mathrm{H}}\right)^{\mathrm{T}}\otimes\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}\right.\\&\quad\left.+\left(\boldsymbol{\theta}^{+}\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\right)^{\mathrm{T}}\otimes\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}\mathbf{R}_{x,t}\mathbf{P}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}^{\perp}\right). \quad (59)\end{aligned}$$

Using the relation between the Kronecker product and the vectorization operator in [51, Lemma 2.11], (59) simplifies to

$$\begin{aligned}\mathcal{D}_{\boldsymbol{\theta}}J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta})&=-\boldsymbol{\theta}^{+}\mathbf{P}_{\boldsymbol{\theta}\mathbf{G}_{\bar{K}}}^{\angle}\mathbf{R}_{x,t}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle\mathrm{H}}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}\\&\quad-\boldsymbol{\theta}^{+}\mathbf{G}_{\bar{K}}\left(\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\boldsymbol{\theta}}^{\perp}\mathbf{G}_{\bar{K}}\right)^{-1}\mathbf{G}_{\bar{K}}^{\mathrm{H}}\mathbf{P}_{\mathbf{G}_{\bar{K}}\boldsymbol{\theta}}^{\angle}\mathbf{R}_{x,t}\mathbf{P}_{[\mathbf{G}_{\bar{K}},\boldsymbol{\theta}]}^{\perp}. \quad (60)\end{aligned}$$

Inserting (60) into $\nabla_{\boldsymbol{\theta}}J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta})=(\mathcal{D}_{\boldsymbol{\theta}}J_{\mathbf{G}_{\bar{K}}}(\boldsymbol{\theta}))^{\mathrm{H}}$ yields the vector gradient of the cost function given in (20).

REFERENCES

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

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

[3] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[4] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[5] S. Braun, W. Zhou, and E. A. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, USA, 2015, pp. 1–5.

[6] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Bias-compensated informed sound source localization using relative transfer functions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1275–1289, Jul. 2018.

[7] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoustical Soc. America*, vol. 152, no. 1, pp. 107–151, 2022.

[8] D. Fejgin and S. Doclo, "Coherence-based frequency subset selection for binaural RTF-vector-based direction of arrival estimation for multiple speakers," in *Proc. Int. Workshop Acoustic Signal Enhancement*, Bamberg, Germany, 2022, pp. 1–5.

[9] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, Jun. 2019.

[10] H. Gode and S. Doclo, "Adaptive dereverberation, noise and interferer reduction using sparse weighted linearly constrained minimum power beamforming," in *Proc. Eur. Signal Process. Conf.*, Belgrade, Serbia, 2022, pp. 95–99.

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

[12] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[13] M. Tammen, I. Kodrasi, and S. Doclo, "Joint estimation of RETF vector and power spectral densities for speech enhancement based on alternating

least squares," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Brighton, U.K., 2019, pp. 795–799.

[14] C. Li, J. Martinez, and R. C. Hendriks, "Joint maximum likelihood estimation of microphone array parameters for a reverberant single source scenario," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 695–705, 2023.

[15] G. Bologni, R. C. Hendriks, and R. Heusdens, "Wideband relative transfer function (RTF) estimation exploiting frequency correlations," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 731–747, 2025.

[16] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

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

[18] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[19] R. Varzandeh, M. Taseska, and E. A. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. Hands-free Speech Commun. Microphone Arrays*, San Francisco, USA, 2017, pp. 11–15.

[20] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function," in *Proc. Eur. Signal Process. Conf.*, Rome, Italy, 2018, pp. 2499–2503.

[21] A. Brendel, J. Zeitler, and W. Kellermann, "Manifold learning-supported estimation of relative transfer functions for spatial filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Singapore, 2022, pp. 8792–8796.

[22] D. Levi, A. Sofer, and S. Gannot, "peerRTF: Robust MVDR beamforming using graph convolutional network," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 1349–1363, 2025.

[23] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Shanghai, China, 2016, pp. 196–200.

[24] A. Deleforge, S. Gannot, and W. Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *Proc. Eur. Signal Process. Conf.*, Nice, France, 2015, pp. 419–423.

[25] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, Jul. 2019.

[26] C. Li and R. C. Hendriks, "Multimicrophone signal parameter estimation in a multi-source noisy reverberant scenario," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 33, pp. 678–692, 2025.

[27] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal procrustes problem," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 755–769, 2020.

[28] S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 320–332, Feb. 2017.

[29] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Montreal, QC, Canada, 2004, pp. iii–889.

[30] B. Schwartz, S. Gannot, and E. A. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 11, pp. 2209–2222, Nov. 2017.

[31] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.

[32] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. CHiME5 Workshop*, Hyderabad, India, 2018, pp. 35–40.

[33] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[34] S. Horiguchi, Y. Fujita, and K. Nagamatsu, "Block-online guided source separation," in *Proc. IEEE Spoken Lang. Technol. Workshop*, Shenzhen, China, 2021, pp. 236–242.

[35] T. Nakashima and N. Ono, "Inverse-free online independent vector analysis with flexible iterative source steering," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Chiang Mai, Thailand, 2022, pp. 749–753.

[36] H. Gode and S. Doclo, "Covariance blocking and whitening method for successive relative transfer function vector estimation in multi-speaker scenarios," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2023, pp. 1–5.

[37] D. Cherkassky and S. Gannot, "Successive relative transfer function identification using blind oblique projection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 474–486, 2020.

[38] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, Oct. 2013.

[39] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1079–1093, Jun. 2016.

[40] S. Hafezi, A. H. Moore, and P. A. Naylor, "Spatial consistency for multiple source direction-of-arrival estimation and source counting," *J. Acoustical Soc. America*, vol. 146, no. 6, pp. 4592–4603, 2019.

[41] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. Habets, "CountNet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 268–282, Feb. 2019.

[42] S. R. Chetupalli and E. A. P. Habets, "Speaker counting and separation from single-channel noisy mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1681–1692, 2023.

[43] M. Yousefi and J. H. Hansen, "Real-time speaker counting in a cocktail party scenario using attention-guided convolutional neural network," 2021, *arXiv:2111.00316*.

[44] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Overlapped speech detection and speaker counting using distant microphone arrays," *Comput. Speech Lang.*, vol. 72, 2022, Art. no. 101306.

[45] H. Gode and S. Doclo, "DNN-based online source counting based on spatial generalized magnitude squared coherence," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2026, pp. 22692–22696.

[46] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[47] D. Fejgin, W. Middelberg, and S. Doclo, "BRUDEX database: Binaural room impulse responses with uniformly distributed external microphones," in *Proc. ITG Conf. Speech Commun.*, Aachen, Germany, 2023, pp. 126–130, doi: 10.5281/zenodo.7986446.

[48] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, South Brisbane, QLD, Australia, 2015, pp. 5206–5210.

[49] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[50] N. Gößling, D. Marquardt, I. Merks, T. Zhang, and S. Doclo, "Optimal binaural LCMV beamforming in complex acoustic scenarios: Theoretical and practical insights," in *Proc. Int. Workshop Acoustic Signal Enhancement*, Tokyo, Japan, 2018, pp. 381–385.

[51] A. Hjørungnes, *Complex-Valued Matrix Derivatives: With Applications in Signal Processing and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

Henri Gode (Graduate Student Member, IEEE) received the B.Eng. and M.Sc. degrees in engineering physics from the Carl von Ossietzky University of Oldenburg, Oldenburg, Germany, in 2019 and 2020, respectively. Since 2021, he has been working toward the Dr.-Ing. degree with Signal Processing Group, Carl von Ossietzky University of Oldenburg, focusing on speech enhancement in multi-channel and multi-source scenarios.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from KU Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with KU Leuven and McMaster University, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors, Leuven. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and Scientific Advisor for the Branch Hearing, Speech and Audio Technology HSA, Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany. His research interests include signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. He was the recipient of the several best paper awards International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019, and ITG Conference on Speech Communication 2025. He served as Associate Editor for *EURASIP Journal on Advances in Signal Processing* (2014-2019) and Senior Area Editor of IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2021-2025). He was member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing. He is currently a member of the EAA Technical Committee on Audio Signal Processing.