Wideband Relative Transfer Function (RTF) Estimation Exploiting Frequency Correlations

Giovanni Bologni, Richard C. Hendriks, *Senior Member, IEEE*, and Richard Heusdens, *Senior Member, IEEE*

***Abstract*—This article focuses on estimating relative transfer functions (RTFs) for beamforming applications. Traditional methods often assume that spectra are uncorrelated, an assumption that is often violated in practical scenarios due to factors such as time-domain windowing or the non-stationary nature of signals, as observed in speech. To overcome these limitations, we propose an RTF estimation technique that leverages spectral and spatial correlations through subspace analysis. Additionally, we derive Cramér–Rao bounds (CRBs) for the RTF estimation task, providing theoretical insights into the achievable estimation accuracy. These bounds reveal that channel estimation can be performed more accurately if the noise or the target signal exhibits spectral correlations. Experiments with both real and synthetic data show that our technique outperforms the narrowband maximum-likelihood estimator, known as covariance whitening (CW), when the target exhibits spectral correlations. Although the proposed algorithm generally achieves accuracy close to the theoretical bound, there is potential for further improvement, especially in scenarios with highly spectrally correlated noise. While channel estimation has various applications, we demonstrate the method using a minimum variance distortionless (MVDR) beamformer for multichannel speech enhancement. A free Python implementation is also provided.**

***Index Terms*—Acoustic parameter estimation, CRB, channel, correlation, Cramér–Rao bound, RTF, relative transfer function.**

I. Introduction

SPATIAL filtering techniques can extract a target signal from the measurements of multiple sensors, also referred to as *beamforming* [1], [2]. Most beamforming techniques, such as the minimum variance distortionless beamformer (MVDR), rely on the knowledge of the relative transfer function (RTF) between a target emitter and a sensor array to virtually *steer* the array towards the direction of interest [3], [4]. RTFs generalize the angle or direction-of-arrival (DOA) concept in scenarios involving the proximity of the source to the receivers or the presence of reflections. These scenarios commonly arise in acoustics and wireless communications, radar and sonar sensing, seismology, and medical imaging.

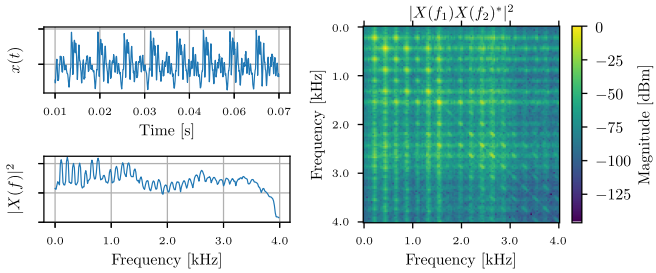

Fig. 1. The /ä/ phoneme uttered by a male speaker. The top left plot depicts the waveform, while the bottom left plot shows the power spectral density (PSD). The peaks in the PSD are found at integer multiples of the fundamental frequency (harmonics). The right plot shows the spectral correlation or bifrequency spectrum. The grid-like structure of peaks in the bifrequency spectrum, whose spacing is proportional to the fundamental frequency, indicates a correlation between harmonic components [12].

One fundamental assumption shared among many channel estimation techniques is that RTFs can be estimated independently per frequency bin after transforming the received signal to the short-time Fourier transform (STFT) domain [5], [6], [7], [8], [9], [10]. This implies that the signals are realizations of wide-sense stationary (WSS) processes or that distinct frequency components of the signal are mutually uncorrelated. It was shown that distinct frequency components of a random process are statistically uncorrelated if and only if the process is WSS [11].

However, the spectral uncorrelation assumption is frequently violated in practice. The STFT coefficients of the signals in neighboring frequency bands are correlated due to the use of short frame lengths and overlap-add/save techniques. In wireless communications, non-stationarity might be due to natural phenomena like the Doppler effect or artificial manipulations such as in orthogonal frequency division multiplexing (OFDM) [12], [13]. In the audio processing domain, vowels are often modeled as an impulse train filtered by a time-varying linear filter. Fig. 1 shows the waveform $x(t)$ of the /ä/ phoneme uttered by a male speaker, its power spectral density (PSD), and its bifrequency spectrum. The bifrequency spectrum approximates $\mathbb{E}[X(f_1)X(f_2)^*]$ for all frequencies f_1, f_2, where $\mathbb{E}[\cdot]$ indicates the expected value and $X(f)$ is the Fourier transform of $x(t)$. The vowel in Fig. 1 has a non-diagonal bifrequency spectrum, implying that its frequency components are correlated. First of all, this is not in line with the typical assumptions being made:

Received 15 December 2023; revised 26 August 2024; accepted 13 January 2025. Date of publication 24 January 2025; date of current version 10 February 2025. This work was supported in part by Dutch Research Council (NWO) and in part by the Signal Processing Systems Group, Delft University of Technology, Delft, The Netherlands. The associate editor coordinating the review of this article and approving it for publication was Dr. Ina Kodrasi. *(Corresponding author: Giovanni Bologni.)*

Giovanni Bologni and Richard C. Hendriks are with the Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: G.Bologni@tudelft.nl; R.C.Hendriks@tudelft.nl).

Richard Heusdens is with the Faculty of Military Sciences, Netherlands Defence Academy (NLDA), 1781 AC Den Helder, The Netherlands (e-mail: R.Heusdens@tudelft.nl).

Digital Object Identifier 10.1109/TASLPRO.2025.3533371

estimation of parameters or processes from such an acoustic scene could be impaired. Secondly, we can conclude that $x(t)$ cannot be modeled as a realization of a WSS process, and the ergodicity assumption does not hold [14], [15]. Characterizing the spectral covariance of such a process requires a phase-adjusted estimator, whose details are discussed in this contribution as well.

Empirical studies on human auditory perception consistently highlight the practical importance of spectral correlations in spatial filtering. These correlations are critical in tasks such as sound localization and speech intelligibility. For instance, speech intelligibility in noise is influenced by the periodic structure of signals, with harmonically complex tones allowing for easier detection compared to inharmonic noise [16]. Additionally, humans can localize speakers based on spatially aliased measurements, but only when spectrally complex sounds are present [17], [18]. Dmochowsky et al. proved that spatial aliasing, a common issue in narrowband signals [19], has reduced impact when the signals are wideband, regardless of the spatial sampling period [20]. Despite the compelling evidence of the relevance of wideband patterns, traditional channel estimation algorithms have rarely considered them explicitly.

Therefore, this paper aims to investigate the impact of spectral correlations on the channel estimation task. Our contributions are twofold: Firstly, we propose an RTF estimation technique based on subspace analysis that exploits spectral and spatial correlations. This technique consistently outperforms the narrowband maximum-likelihood estimator (MLE), known as covariance whitening (CW) [21], [22], [23], [24], when the target exhibits spectral correlations. Secondly, we derive conditional and unconditional CRBs for the RTF estimation task. To the best of our knowledge, bounds for the RTF estimation task have not been derived before, not even for the narrowband scenario. The bounds show that channel estimation can be conducted more accurately if the target or the additive noise presents inter-frequency correlations. Our findings align with experiments showing that both parametric methods and methods based on deep neural network (DNN) for speech enhancement, which jointly process spectral information, outperform their counterparts that process each frequency bin independently [25], [26], [27], [28]. Although the accuracy of the proposed algorithm is generally close to the bound, there is some room for improvement, especially when noise signals with high spectral correlation are present. An additional contribution is that, in the spirit of reproducible research, a Python implementation is freely available online.[1]

The article details the signal model in Section II. In Section III, we demonstrate how to recover the spectral-spatial covariance matrix of the source at the receivers, and introduce two related RTF estimation methods. Based on these results, we propose a novel algorithm for RTF estimation in Section IV. To better assess the algorithms' performance, we compare them to the lower bounds on the variance of RTF estimation, which are derived in Section V. Numerical evidence of the superiority of the proposed algorithm, especially when the target presents spectral correlation, is provided in Section VI. In Section VII, we present additional discussion and insights on the experiments. Finally, some conclusions are drawn in Section VIII, summarizing the essential findings and contributions of this paper.

[1]https://github.com/Screeen/SVD-direct

II. Signal Model

In a reverberant and noisy environment, we consider the case of a single point source impinging on an array of $M \geq 2$ sensors. The signal received by the array is given in the STFT domain as:

$$\boldsymbol{x}_k(l) = \boldsymbol{d}_k(l) + \boldsymbol{v}_k(l) = s_k(l)\boldsymbol{a}_k + \boldsymbol{v}_k(l) \in \mathbb{C}^M, \quad (1)$$

where $\boldsymbol{d}_k(l) = s_k(l)\boldsymbol{a}_k$ is the target signal at the receiver, $l = 1, \ldots, L$ is the time-frame index and the subscript $k = 1, \ldots, K$ denotes the frequency bin index. The STFT coefficients of the target signal at the source are modeled by $s_k(l)$, which are realizations of complex random variables with zero mean. The target coefficients are *not* assumed to be mutually independent over frequency. They can follow any probability distribution. The transfer function $\boldsymbol{a}_k \in \mathbb{C}^M$ models the wave propagation from the target point source to the M sensors. The transfer function is assumed to be an unknown deterministic quantity that typically needs to be estimated in beamforming applications. The noise coefficients $\boldsymbol{v}_k(l)$ are also modeled as complex random variables with zero mean and an arbitrary probability distribution.

Let us now consider the coefficients for all frequency components jointly. Noisy coefficients corresponding to a single time frame l, for M sensors, at K frequencies, can be stacked in a column vector as in $\boldsymbol{x} = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_K^T]^T \in \mathbb{C}^{KM}$. The time-frame index l is left out for notational convenience. In a similar fashion, noise vectors $\boldsymbol{v}_k$, transfer function vectors $\boldsymbol{a}_k$ and desired signal $\boldsymbol{d}_k$ can be stacked vertically to form $\boldsymbol{v}$, $\boldsymbol{a}$, and $\boldsymbol{d}$, respectively, so that $\boldsymbol{x} = \boldsymbol{d} + \boldsymbol{v}$. In this case, it is helpful to collect the signal coefficients s_k in a random vector $\bar{\boldsymbol{s}} = [s_1, s_2, \ldots, s_K]^T$. Let us also define $\boldsymbol{s} = \bar{\boldsymbol{s}} \otimes \mathbf{1}_M = [s_1\mathbf{1}_M^T, s_2\mathbf{1}_M^T, \ldots, s_K\mathbf{1}_M^T]^T$, where $\otimes$ is the Kronecker product and $\mathbf{1}_M$ is the M-dimensional all-ones vector. Next, let

$$\boldsymbol{A} = \mathrm{diag}(\boldsymbol{a}) = \mathrm{diag}(a_{11}, \ldots, a_{1M}, a_{21}, \ldots a_{KM}), \quad (2)$$

contain the transfer functions for all frequencies and sensors. The vector of desired signals is then given by

$$\boldsymbol{d} = \boldsymbol{A}\boldsymbol{s} = \boldsymbol{A}(\bar{\boldsymbol{s}} \otimes \mathbf{1}_M), \quad (3)$$

such that the noisy coefficients for the wideband model can be written as

$$\boldsymbol{x} = \boldsymbol{d} + \boldsymbol{v} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{v}. \quad (4)$$

Notice that (4) generalizes the narrowband model with multiplicative transfer function (MTF) approximation (1) to a wideband scenario. In the MTF approximation, the linear convolution in the time domain is represented as multiplication in the STFT domain [1]. This constrains the transfer functions $\boldsymbol{a}_k$ to be at most K samples long in the time domain, effectively capturing the early reflections only, and neglecting the late reverberation components.

Next, we model the spatial and spectral correlations between the signals. Spatial correlation matrices are widely used in array processing to model relations between signals received at different sensors. Here, we also consider *spectral* correlations between different frequency components. The spectral-spatial covariance matrix $\boldsymbol{R}_x = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H] \in \mathbb{C}^{KM\times KM}$, can be expressed as

$$\boldsymbol{R}_x = \begin{bmatrix} \boldsymbol{r}_x(1,1) & \boldsymbol{r}_x(1,2) & \cdots & \boldsymbol{r}_x(1,K) \\ \boldsymbol{r}_x(2,1) & \boldsymbol{r}_x(2,2) & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{r}_x(K,1) & \boldsymbol{r}_x(K,2) & \cdots & \boldsymbol{r}_x(K,K) \end{bmatrix}, \quad (5)$$

where $(\cdot)^H$ indicates the conjugate transpose operation, and $\boldsymbol{r}_x(i,j) = \mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_j^H] \in \mathbb{C}^{M\times M}$ is the spectral-spatial covariance matrix at two arbitrary frequencies i, j. When noise and target signal are statistically uncorrelated, we have $\boldsymbol{R}_x = \boldsymbol{R}_d + \boldsymbol{R}_{\boldsymbol{v}}$, that is, $\boldsymbol{r}_x(i,j) = \mathbb{E}[s_i s_j^*]\boldsymbol{a}_i\boldsymbol{a}_j^H + \mathbb{E}[\boldsymbol{v}_i\boldsymbol{v}_j^H]$. Let us now introduce alternative formulations of the covariance matrices that will be useful for our analysis. Using the definition in (3), the signal covariance matrix $\boldsymbol{R}_d = \mathbb{E}[\boldsymbol{d}\boldsymbol{d}^H]$ can be expressed as

$$\boldsymbol{R}_d = \mathbb{E}[\boldsymbol{A}\boldsymbol{s}\boldsymbol{s}^H\boldsymbol{A}^H] = \boldsymbol{A}\,\mathbb{E}[\boldsymbol{s}\boldsymbol{s}^H]\boldsymbol{A}^H = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H, \quad (6)$$

where $\boldsymbol{R}_s$ is defined as $\boldsymbol{R}_s = \mathbb{E}[\boldsymbol{s}\boldsymbol{s}^H]$. Using the properties of the Kronecker product, the covariance matrix $\boldsymbol{R}_s$ can, in turn, be rewritten as

$$\begin{aligned} \boldsymbol{R}_s &= \mathbb{E}[(\bar{\boldsymbol{s}} \otimes \mathbf{1}_M)(\bar{\boldsymbol{s}} \otimes \mathbf{1}_M)^H] = \mathbb{E}[\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}^H \otimes \mathbf{1}_M\mathbf{1}_M^H] \\ &= \mathbb{E}[\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}^H] \otimes \mathbf{1}_{M\times M} = \boldsymbol{R}_{\bar{s}} \otimes \mathbf{1}_{M\times M}, \end{aligned} \quad (7)$$

where $\mathbf{1}_{M\times M}$ is the all-ones matrix of size $M \times M$ and

$$\boldsymbol{R}_{\bar{s}} = \mathbb{E}[\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}^H] \in \mathbb{C}^{K\times K}. \quad (8)$$

III. Background Information

This section begins by reporting a strategy to estimate sample spectral-spatial covariance matrices. It then demonstrates that the desired signal covariance matrix $\boldsymbol{R}_d$ is singular, with its rank being limited by the number of frequency components K. Section III-C explores how the eigenvectors of $\boldsymbol{R}_x$ and $\boldsymbol{R}_d$ are affected by additive noise, and it reports a strategy for recovering $\boldsymbol{R}_d$. Finally, two well-known algorithms for RTF estimation are introduced.

A. Phase-Adjusted Sample Covariance Matrix

The commonly used sample covariance matrix estimate, serving as the MLE for jointly Gaussian WSS data, is expressed as

$$\tilde{\boldsymbol{R}}_x = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{x}(l)\boldsymbol{x}(l)^H, \quad (9)$$

where l is the realization index. Alternatively, l can be treated as a time-frame index assuming second-order ergodicity.

However, when spectral correlations are present, the WSS assumption becomes inaccurate, requiring an alternative estimator for the spectral-spatial covariance matrices. In the estimation of *spectral* correlations from STFT data, it is crucial to establish a connection among phase components across all frames and frequencies. In most implementations of the STFT, phase components are linked to the beginning of each frame. Therefore, there is a need to connect these phase components to a common reference point, such as the signal's onset, as mentioned by Antoni [29]. The phase-adjusted noisy STFT data at frequency k is given by:

$$\mathbf{x}_k^c(l) = \mathbf{x}_k(l)e^{-j2\pi lRk/K}, \quad l = 1,\ldots,L, \quad (10)$$

where R denotes the block shift between frames.

Let us examine the impact of phase correction through an example. Consider a harmonic signal of the form $y(t) = \sum_{h=1}^{3}\cos(2\pi f_0 ht)$, $t \in \mathbb{N}$, where f_0 is the fundamental frequency in normalized units, and h denotes the harmonic index. The harmonic components at frequencies $f_0 h$ for $h = 1, 2, 3$ are deterministic, thus perfectly correlated, meaning that knowing one component allows us to infer the value of another. In the STFT domain, we denote the harmonic signal as $y_k(l)$, $l = 1,\ldots,L$, and its phase-corrected version as $y_k^c(l)$. The overlap of the STFT is set to 75%, corresponding to $R = K/4$. Fig. 2(a) shows the phase components of the three non-zero frequency components of $y_k(l)$ across time frames. Due to the misalignment between the block-shift R and the periodicities of $y(t)$, the phases of the harmonics components appear to change randomly from frame to frame. However, after applying phase correction to get $y_k^c(l)$, we can accurately determine the phase of all components (Fig. 2(b)) with respect to the time origin, $t = 0$. Let us also analyze the impact of phase correction on the estimation of the spectral correlations. For the phase-corrected signal $y_k^c(l)$ (Fig. 2(d)), the spectral correlation is maximal across all components, while the original $y_k(l)$ signal incorrectly appears to exhibit a lower spectral correlation due to spurious effects of phase cancellation (Fig. 2(c)).

The phase correction becomes superfluous when dealing with products of components at the same frequency, as the conjugation leads to the cancellation of the phase term: $\mathbf{x}_k^c(l)\mathbf{x}_k^c(l)^H = \mathbf{x}_k(l)\mathbf{x}_k(l)^H$. Similarly, the exponential term in (10) is identical to one, thus ineffective, when $R = K$, i.e., when adjacent frames do not overlap, or when independent realizations of the signals are used. Therefore, the correction of (10) is applied solely in Sections VI-D and VI-C to the overlapping STFT frames of real speech signals before covariance matrix estimation, so that, for $k_1, k_2 = 1,\ldots,K$,

$$\begin{aligned} \hat{\boldsymbol{r}}_x(k_1,k_2) &= \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{x}_{k_1}^c(l)\boldsymbol{x}_{k_2}^c(l)^H \\ &= \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{x}_{k_1}(l)\boldsymbol{x}_{k_2}(l)^H e^{-j2\pi lR(k_1-k_2)/K}. \end{aligned} \quad (11)$$

B. Upper Bound on the Rank of Target Covariance Matrix

Lemma 1: $\operatorname{rank}(\boldsymbol{R}_d) \le K$

Proof: To support this claim, we first state two well-known properties of the matrix rank. Consider two matrices $\boldsymbol{X} \in \mathbb{C}^{m\times n}$ and $\boldsymbol{Y} \in \mathbb{C}^{n\times p}$. According to [30], we have that:

$$\operatorname{rank}(\boldsymbol{X}\boldsymbol{Y}) \le \min(\operatorname{rank}(\boldsymbol{X}), \operatorname{rank}(\boldsymbol{Y})), \quad (12)$$

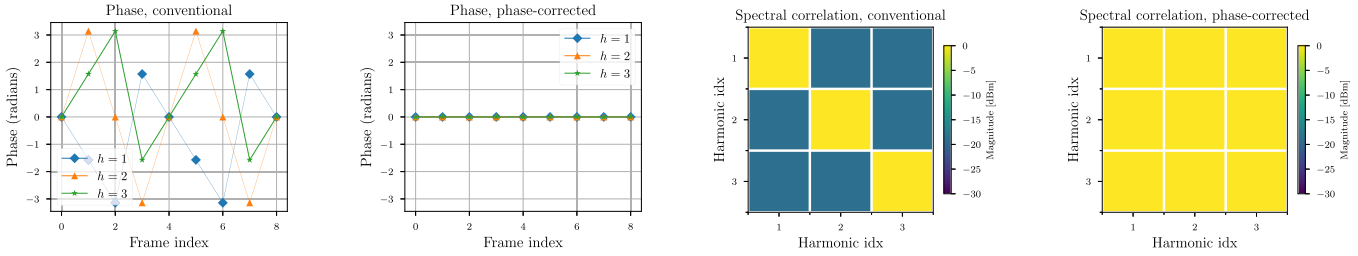

Fig. 2. Phase correction improves estimation of the spectral covariance. (a) Phase components of the STFT of the original signal $y_k(l)$; (b) spectral covariance of $y_k(l)$; (c) phase components of the STFT of the phase-corrected signal $y_k^c(l)$; (d) spectral covariance of $y_k^c(l)$.

$$\operatorname{rank}(\boldsymbol{X} \otimes \boldsymbol{Y}) = \operatorname{rank}(\boldsymbol{X}) \operatorname{rank}(\boldsymbol{Y}). \tag{13}$$

The covariance matrix $\boldsymbol{R}_{\bar{s}} = \mathbb{E}[\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}^H]$ in (8) obeys $\operatorname{rank}(\boldsymbol{R}_{\bar{s}}) \leq K$. The rank of the all-one matrix, instead, is $\operatorname{rank}(\mathbf{1}_{M\times M}) = 1$. From (7) and the rank property of Kronecker products in (13) it follows that

$$\operatorname{rank}(\boldsymbol{R}_s) = \operatorname{rank}(\boldsymbol{R}_{\bar{s}} \otimes \mathbf{1}_{M\times M}) = \operatorname{rank}(\boldsymbol{R}_{\bar{s}}) \leq K. \tag{14}$$

Moreover, let at least K of the coefficients of the diagonal RTF matrix $\boldsymbol{A}$ be non-zero by assumption, so that $\operatorname{rank}(\boldsymbol{A}) \geq K$. It is now possible to analyze the matrix rank of $\boldsymbol{R}_d$:

$$\operatorname{rank}(\boldsymbol{R}_d) = \operatorname{rank}(\boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H) \tag{15}$$

$$\leq \min\left(\operatorname{rank}(\boldsymbol{A}), \operatorname{rank}(\boldsymbol{R}_s)\right) \leq K, \tag{16}$$

where the inequality follows from the rank matrix product property in (12) and (14). This completes the proof. □

C. Estimation of the Target Covariance Matrix

Suppose that $\boldsymbol{R}_x$ is known, and let γ^2 be the noise variance. Estimated quantities are denoted as $\hat{(\cdot)}$. For example, the estimated noise variance is represented as $\hat{\gamma}^2$. Assuming that the noise exhibits uniform power across both space and frequency, remaining uncorrelated in both domains, we have $\boldsymbol{R}_v = \gamma^2 \boldsymbol{I}_{KM}$. As the identity matrix is diagonalizable by any unitary matrix,

$$\boldsymbol{R}_x = \boldsymbol{R}_d + \gamma^2\boldsymbol{I} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^H + \gamma^2\boldsymbol{I} = \boldsymbol{V}(\boldsymbol{\Lambda} + \gamma^2\boldsymbol{I})\boldsymbol{V}^H,$$

where $\boldsymbol{V}$ is the eigenvector matrix of $\boldsymbol{R}_d$, and $\boldsymbol{\Lambda}$ is the diagonal matrix containing the eigenvalues of $\boldsymbol{R}_d$. Therefore, if the estimated, phase-adjusted sample covariance matrix is decomposed as $\hat{\boldsymbol{R}}_x = \hat{\boldsymbol{V}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{V}}^H$, the covariance matrix of the target at the sensors can be approximated by $\hat{\boldsymbol{R}}_d = \hat{\boldsymbol{V}} \max(\hat{\boldsymbol{\Lambda}} - \hat{\gamma}^2\boldsymbol{I}, 0)\hat{\boldsymbol{V}}^H$, where the $\max(\cdot,\cdot)$ operator forces the eigenvalues of the Hermitian positive semidefinite (HPSD) matrix $\hat{\boldsymbol{R}}_d$ to be non-negative.

If spatially or spectrally colored noise is present, the eigenvectors of $\boldsymbol{R}_x$ and $\boldsymbol{R}_d$ will differ. However, estimating $\hat{\boldsymbol{R}}_d$ and computing its eigenvalue decomposition is still possible if an estimate of the noise covariance matrix $\hat{\boldsymbol{R}}_v$ is available and it is full-rank, hence invertible. To ensure that this requirement is satisfied, we apply *diagonal loading*, which consists of adding a scaled identity matrix to the estimated noise covariance matrix: $\hat{\boldsymbol{R}}_v \leftarrow \hat{\boldsymbol{R}}_v + \epsilon\boldsymbol{I}$, where ϵ is a small positive value. The clean covariance matrix $\hat{\boldsymbol{R}}_d$ can be estimated from the generalized eigenvalue decomposition (GEVD) of $\hat{\boldsymbol{R}}_x$ and $\hat{\boldsymbol{R}}_v$ or from the eigenvalue decomposition of the prewhitened noisy covariance matrix $\hat{\boldsymbol{R}}_v^{-1/2}\hat{\boldsymbol{R}}_x\hat{\boldsymbol{R}}_v^{-1/2}$. The present examination will be limited to the GEVD because the two procedures are theoretically equivalent [31], [32].[2] Given the estimates $\hat{\boldsymbol{R}}_x$ and $\hat{\boldsymbol{R}}_v$, an estimate of the desired covariance matrix $\hat{\boldsymbol{R}}_d$ can be obtained as follows:

1) Computation of $\hat{\boldsymbol{R}}_x\boldsymbol{U} = \hat{\boldsymbol{R}}_v\boldsymbol{U}\boldsymbol{D}$ or, equivalently, $\boldsymbol{Q}^H\hat{\boldsymbol{R}}_x = \boldsymbol{D}\boldsymbol{Q}^H\hat{\boldsymbol{R}}_v$, where $\boldsymbol{D}$ are the generalized eigenvalues, $\boldsymbol{U}$ are the right generalized eigenvectors, $\boldsymbol{Q}$ are the left generalized eigenvectors, and $\boldsymbol{U} = \boldsymbol{Q}^{-H}$.
2) Partitioning of the left eigenvectors $\boldsymbol{Q} = [\boldsymbol{Q}_x\, \boldsymbol{Q}_v]$, where $\boldsymbol{Q}_x$ comprises of the first K_d columns of $\boldsymbol{Q}$.
3) Estimation of $\hat{\boldsymbol{R}}_d$ as $\hat{\boldsymbol{R}}_d = \boldsymbol{Q}_x \max(\boldsymbol{D}_x - \boldsymbol{I}, 0)\boldsymbol{Q}_x^H$,

where K_d is the estimated rank of $\boldsymbol{R}_d$, and $\boldsymbol{D}_x$ is a diagonal subblock formed by the first K_d columns and rows of $\boldsymbol{D}$. By virtue of Lemma 1, $K_d \leq K$. The number of frames L available for estimating the covariance matrices also constrains the maximum possible matrix rank, such that $K_d \leq L$. As a consequence, in steps 2) and 3), $K_d = \min(K, L)$ eigenvalue-eigenvector pairs are retained. Note that, due to the sparse spectral distribution of speech, the actual rank of $\boldsymbol{R}_d$ might be lower than K_d. Specifically, since many frequency components of speech signals are zero, the corresponding rows and columns in $\boldsymbol{R}_d$ will also be zero, reducing the rank of the matrix.

D. Covariance Whitening and Covariance Subtraction

The GEVD routine detailed in Section III-C is also widely used in traditional, narrowband processing for estimating the target spatial covariance matrix. It is indeed at the core of the covariance whitening (CW) algorithm, one of the most effective techniques for RTF estimation [21], [22], [23], [24]. Let the (narrowband) noisy spatial covariance matrix be represented by $\boldsymbol{R}_x(k) = \mathbb{E}[\boldsymbol{x}_k(l)\boldsymbol{x}_k(l)^H] \in \mathbb{C}^{M\times M}$ and the noise spatial covariance matrix by $\boldsymbol{R}_v(k)$. The CW technique consists of estimating the generalized left eigenvectors of $(\boldsymbol{R}_x(k), \boldsymbol{R}_v(k))$

[2] A standard routine for computing the GEVD of HPSD matrices is based on Cholesky decomposition [30, Algorithm 8.7.1]. It is used in the popular LAPACK drivers [33] that are the backbone of Matlab and Numpy/Scipy.

for each discrete frequency k, and then retaining the eigenvector corresponding to the largest eigenvalue. Assuming that a single speaker is present, the rank of $\boldsymbol{R}_d(k) = \mathbb{E}[|s_k|^2]\boldsymbol{a}_k\boldsymbol{a}_k^H$ is 1. Therefore, the principal eigenvector equals the target RTF $\boldsymbol{a}_k$ up to a multiplicative factor.

Covariance subtraction (CS) is another popular technique for RTF estimation. CS estimates the target spatial covariance matrix by subtracting the noise covariance matrix from the observed covariance matrix, i.e., $\boldsymbol{R}_d^{\text{CS}}(k) = \boldsymbol{R}_x(k) - \boldsymbol{R}_v(k)$. The RTF is then estimated from the principal eigenvector of $\boldsymbol{R}_d^{\text{CS}}(k)$. This simpler technique is generally less accurate then CW [23].

IV. Proposed RTF Estimation Algorithm: SVD-Direct

In the preceding sections, the investigation focused on the spectral-spatial covariance matrix of a noisy signal received from multiple sensors. The knowledge gained from this investigation can be applied to estimate the channel $\boldsymbol{a}$, provided that estimates of the spectral-spatial covariance for both the noisy signal $\hat{\boldsymbol{R}}_x$ and the noise-only signal $\hat{\boldsymbol{R}}_v$ are available. To this aim, we introduce a new method for RTF estimation. The proposed algorithm is based on a row partitioning of the estimated spectral-spatial covariance $\hat{\boldsymbol{R}}_d$, followed by an SVD on each frequency subblock. The approach is named **SVD-direct** to emphasize the simplicity of its implementation and the central role played by the singular value decomposition. The proposed method extends the CW technique (Section III-D) to a wideband scenario, thus leveraging inter-frequency correlations for better estimation accuracy. Unlike CW, multiple frequency components are processed simultaneously both in the prewhitening and in the ensuing decomposition step.

The basic idea of the proposed RTF estimation algorithm can be explained by an example. First, let us introduce a simplified case with $K = 2$ frequency components, to gain some intuition on the structure of $\boldsymbol{R}_d = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H$. We have that

$$\boldsymbol{R}_d = \begin{bmatrix} \mathbb{E}[|s_1|^2]\boldsymbol{a}_1\boldsymbol{a}_1^H & \mathbb{E}[s_1 s_2^*]\boldsymbol{a}_1\boldsymbol{a}_2^H \\ \mathbb{E}[s_2 s_1^*]\boldsymbol{a}_2\boldsymbol{a}_1^H & \mathbb{E}[|s_2|^2]\boldsymbol{a}_2\boldsymbol{a}_2^H \end{bmatrix} = \tag{17}$$

$$= \begin{bmatrix} \sigma_1^2\boldsymbol{a}_1\boldsymbol{a}_1^H & \sigma_{12}\boldsymbol{a}_1\boldsymbol{a}_2^H \\ \sigma_{12}^*\boldsymbol{a}_2\boldsymbol{a}_1^H & \sigma_2^2\boldsymbol{a}_2\boldsymbol{a}_2^H \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_d^{(1)} \\ \boldsymbol{R}_d^{(2)} \end{bmatrix}, \tag{18}$$

where we have introduced the auxiliary variables $\sigma_1^2 = \mathbb{E}[|s_1|^2]$, $\sigma_2^2 = \mathbb{E}[|s_2|^2]$, $\sigma_{12} = \mathbb{E}[s_1 s_2^*]$ to simplify the notation. The transfer function for the ith frequency is $\boldsymbol{a}_i \in \mathbb{C}^M$. We also defined the block-matrices $\boldsymbol{R}_d^{(1)}, \boldsymbol{R}_d^{(2)} \in \mathbb{C}^{M\times 2M}$. The absence of spectral correlations in the source signal $\boldsymbol{s}$ would lead to $\mathbb{E}[s_1 s_2^*] = \mathbb{E}[s_2 s_1^*] = 0$. Now, consider the block matrix $\boldsymbol{R}_d^{(1)} = [\sigma_1^2\boldsymbol{a}_1\boldsymbol{a}_1^H \quad \sigma_{12}\boldsymbol{a}_1\boldsymbol{a}_2^H]$ in (18). Notice that $\boldsymbol{R}_d^{(1)}$ is a rank-1 matrix, whose left principal singular vector is proportional to $\boldsymbol{a}_1$. The right principal singular vector of $\boldsymbol{R}_d^{(1)}$ is proportional to $[\boldsymbol{a}_1^T \ \boldsymbol{a}_2^T]^T$. To see this, consider the matrix product

$$\boldsymbol{R}_d^{(1)}(\boldsymbol{R}_d^{(1)})^H = (\sigma_1^2\|\boldsymbol{a}_1\|^2 + \sigma_{12}^2\|\boldsymbol{a}_2\|^2)\,\boldsymbol{a}_1\boldsymbol{a}_1^H \tag{19}$$

Algorithm 1: SVD-Direct.

Input: $\hat{\boldsymbol{R}}_x, \hat{\boldsymbol{R}}_v, M, K$
Output: RTF estimates $\hat{\boldsymbol{a}}$

```
# Estimate R̂_d from the GEVD (Section III-C).
R̂_d ← GEVD_routine(R̂_x, R̂_v)

# Partition in K "fat" M × KM blocks
[(R̂_d^(1))^T, (R̂_d^(2))^T, ..., (R̂_d^(K))^T]^T ← R̂_d

# Per each frequency
for k = 1, ..., K do
  P^(k) D^(k) Q^(k) ← SVD(R̂_d^(k))

  # Rescale left principal singular vectors
  â^(k) ← Normalize(p_1^(k)).
end for
```

from which it follows that $\boldsymbol{R}_d^{(1)}(\boldsymbol{R}_d^{(1)})^H$ is a rank-1 matrix with principal eigenvector $\boldsymbol{a}_1$.[3] It follows that by decomposing $\boldsymbol{R}_d^{(1)}$ with an SVD and selecting the principal left singular component, $\boldsymbol{a}_1$ can be recovered up to a scalar factor.

The procedure above can be repeated for each subblock $\boldsymbol{R}_d^{(k)} \in \mathbb{C}^{M\times KM}$, $k = 1, \ldots, K$, leading to the proposed wideband channel estimation method, **SVD-direct** (Algorithm 1).

The function `Normalize` is defined as $\texttt{Normalize}(\boldsymbol{a}^{(k)}) = \boldsymbol{a}^{(k)}/[\boldsymbol{a}^{(k)}]_r$, and $[\boldsymbol{a}^{(k)}]_r$ is the entry corresponding to the r-th (reference) sensor.

V. Cramér–Rao Lower Bound

Based on the spectral-spatial covariance matrix of the signal received at the multiple sensors, we derived an algorithm for RTF estimation, taking correlation across frequency into account. To determine how close this algorithm is to the optimal performance, we compare it to the CRB.

In the following, we first define the CRB and show how to derive it when estimating a deterministic function of an unknown parameter. The CRB is then calculated for two scenarios: (i) a setting where the target signal $\boldsymbol{s}(l)$ is deterministic and known (Section V-B), and (ii) a scenario where the target signal has a known covariance matrix $\boldsymbol{R}_s$, but the signal realizations are unknown (Section V-C). Note that the former bound will lead to an unrealistic lower bound, as in the current scenario, $\boldsymbol{s}(l)$ is never known. The latter bound is realistic as it only assumes that the first- and second-order statistics are known. The two settings are also known as the deterministic or conditional CRB, and stochastic or unconditional CRB, respectively [34]. Although the CRBs are derived for the wideband scenario, they encompass the bounds for narrowband RTF estimation as a specific case.

It is worth noting that the CRB for proper complex-valued multivariate Gaussian parameters has been previously explored.

[3] Throughout the paper, $\|\cdot\|$ indicates the 2-norm.

In [35, Eq. 15.52], an approach that treats the real and imaginary components of the parameters independently was adopted. Conversely, in [36, Eq. 6.55], the Wirtinger derivatives were employed. However, neither of these references extends its analysis to incorporate further deterministic transformations.

A. Problem Formulation

Let us consider the case where the parameters $\boldsymbol{\theta}$ to be estimated are complex-valued, deterministic but unknown, and the observed data matrix is $\boldsymbol{X} = [\boldsymbol{x}(1) \ \dots \ \boldsymbol{x}(L)]$. The distribution of the observed data is $p(\boldsymbol{X};\boldsymbol{\theta})$. The Fisher information matrix (FIM) is found as the negative expected Hessian of the log-likelihood function:

$$\boldsymbol{I}_{\boldsymbol{\theta}} = -\mathbb{E}[\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^H \ln p(\boldsymbol{X};\boldsymbol{\theta})] = -\mathbb{E}[\nabla_{\boldsymbol{\theta}}^2 \ln p(\boldsymbol{X};\boldsymbol{\theta})], \quad (20)$$

where the expectation is taken with respect to $p(\boldsymbol{X};\boldsymbol{\theta})$. The gradient and the Hessian are defined as

$$[\nabla_{\boldsymbol{\theta}} f]_i = \partial f / \partial \theta_i, \qquad [\nabla_{\boldsymbol{\theta}}^2 f]_{ij} = \partial^2 f / \partial \theta_i \partial \theta_j^*,$$

and the partial derivatives are Wirtinger derivatives [37]. The covariance matrix $\boldsymbol{R}_{\hat{\boldsymbol{\theta}}}$ of any unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ satisfies $\boldsymbol{R}_{\hat{\boldsymbol{\theta}}} \succeq \boldsymbol{I}_{\boldsymbol{\theta}}^{-1}$.[4] When the quantity to estimate is given by a function $\boldsymbol{\phi} = \boldsymbol{g}(\boldsymbol{\theta})$ of some underlying parameter, the bound follows as [38]

$$\boldsymbol{R}_{\hat{\boldsymbol{\phi}}} \succeq (\nabla_{\boldsymbol{\theta}} \boldsymbol{g}) \boldsymbol{I}_{\boldsymbol{\theta}}^{-1} (\nabla_{\boldsymbol{\theta}}^H \boldsymbol{g}), \quad (21)$$

where $\boldsymbol{R}_{\hat{\boldsymbol{\phi}}}$ is the covariance matrix of the estimator $\hat{\boldsymbol{\phi}} = \boldsymbol{g}(\hat{\boldsymbol{\theta}})$.

In the present case, we define a function $\boldsymbol{g} : \mathbb{C}^{2KM} \mapsto \mathbb{C}^{KM}$ that transforms a transfer function to a *relative* transfer function. It is given by

$$\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{g}([\boldsymbol{a}^T \ \boldsymbol{a}^H]^T) = \boldsymbol{a} / \boldsymbol{a}_{\text{ref}}, \quad (22)$$

where the division is intended element-wise and

$$\boldsymbol{a}_{\text{ref}} = [a_{1r} \boldsymbol{1}_M^T, a_{2r} \boldsymbol{1}_M^T, \dots, a_{Kr} \boldsymbol{1}_M^T]^T,$$

is the vector with the responses of the rth (reference) sensor at all frequencies. Notice that $\boldsymbol{g}(\cdot)$ corresponds to the `Normalize`$(\cdot)$ function defined in Section IV, with the only difference that $\boldsymbol{g}(\cdot)$ acts on transfer functions for all frequencies and sensors simultaneously. This function can be readily modified to accommodate various strategies for reference sensor selection [1, Eq. 10].

B. Conditional Cramér–Rao Bound

Consider the model from (4):

$$\boldsymbol{x}(l) = \boldsymbol{A}\boldsymbol{s}(l) + \boldsymbol{v}(l), \quad l = 1, \dots, L. \quad (23)$$

Firstly, we analyze the case where the signal $\boldsymbol{s}(l)$ is known and the absolute transfer function $\boldsymbol{A}$, defined in (2), is deterministic but unknown. The noise $\boldsymbol{v}(l)$ is a complex circular Gaussian random process with known spectral-spatial covariance $\boldsymbol{R}_v$. The vector of unknown parameters is $\boldsymbol{\theta} = [\boldsymbol{a}^T \boldsymbol{a}^H]^T \in \mathbb{C}^{2KM}$. The observed data $\boldsymbol{X}$ follows a complex Gaussian distribution so that the log-likelihood is given by

$$\ln p(\boldsymbol{X};\boldsymbol{\theta}) = -L \ln |\pi \boldsymbol{R}_v| - \sum_{l=1}^{L} \boldsymbol{v}(l)^H \boldsymbol{R}_v^{-1} \boldsymbol{v}(l). \quad (24)$$

We have the following result.

Theorem 1 (Conditional CRB): The variance of any conditional RTF estimator is lower bounded by:

$$\text{CRB}[\boldsymbol{g}(\hat{\boldsymbol{\theta}})]_i = \left[(\nabla_{\boldsymbol{a}} \boldsymbol{g}) (\boldsymbol{B}^*)^{-1} (\nabla_{\boldsymbol{a}}^H \boldsymbol{g}) \right]_{ii}, \quad (25)$$

for $i = 1, \dots, M$, where the matrix $\boldsymbol{B}$ is defined as $\boldsymbol{B} = \sum_{l=1}^{L} \boldsymbol{S}(l)^H \boldsymbol{R}_v^{-1} \boldsymbol{S}(l)$ and $\boldsymbol{S}(l) = \text{diag}(\boldsymbol{s}(l))$.

Proof: See Appendix A. □

1) Interpretation: For ease of analysis, consider the case where the noise is spatially and spectrally uncorrelated, i.e., $\boldsymbol{R}_v = \gamma^2 \boldsymbol{I}_{KM}$. In this scenario, the i-th element on the diagonal of the Fisher information matrix is given by $[\boldsymbol{I}_\theta]_{ii} = \gamma^{-2} \sum_{l=1}^{L} |s_i(l)|^2$. As the noise variance γ^2 increases, the Fisher information $[I_\theta]_{ii}$ decreases. Conversely, increasing the number of frames L available for estimation results in higher Fisher information, as the quantity $|s_i(l)|^2$ is always non-negative. Thus, the achievable accuracy of the RTF estimation decreases with higher noise power and improves with more time frames.

C. Unconditional Cramér–Rao Bound

Consider again the model in (23). This time, we examine the more realistic scenario where the spectral-spatial covariance of the target signal $\boldsymbol{R}_s$ is known but not the signal itself. The transfer function $\boldsymbol{A}$ is again deterministic but unknown. This bound is then expected to be greater than the one derived in Theorem 1 because the target signal is only known up to its second-order statistics. In this case, the log-likelihood function is given by:

$$\ln p(\boldsymbol{X};\boldsymbol{\theta}) = -L \ln |\pi \boldsymbol{R}_x| - L \, \text{tr}\,(\hat{\boldsymbol{R}}_x \boldsymbol{R}_x^{-1}), \quad (26)$$

where $\boldsymbol{R}_x = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^H + \boldsymbol{R}_v$. We have the following result.

Theorem 2 (Unconditional CRB): In the unconditional settings, the variance of any unbiased RTF estimator is lower bounded by:

$$\text{CRB}[\boldsymbol{g}(\hat{\boldsymbol{\theta}})]_i = \left[(\nabla_{\boldsymbol{a}} \boldsymbol{g}) \boldsymbol{C} (\nabla_{\boldsymbol{a}}^H \boldsymbol{g}) \right]_{ii}, \quad (27)$$

for $i = 1, \dots, M$, where $\boldsymbol{C}$ is obtained by selecting the first KM rows and columns from the inverse FIM $\boldsymbol{I}_{\boldsymbol{\theta}}^{-1}$.

Proof: See Appendix B. □

1) Interpretation: Consider the case where the noise is uncorrelated, i.e., $\boldsymbol{R}_v = \gamma^2 \boldsymbol{I}_{KM}$. The i-th element on the diagonal of the Fisher information matrix is given by $[\boldsymbol{I}_\theta]_{ii} = L \, \text{tr}\,(\boldsymbol{R}_x^{-1} \boldsymbol{F}_i \boldsymbol{R}_x^{-1} \boldsymbol{G}_i)$. As the number of frames L available for estimation increases, the Fisher information increases linearly. Thus, as we have seen for the conditional CRB in Section V-B, the achievable accuracy of the RTF estimation improves with more time frames. Also, as the noise variance γ^2 increases, the Fisher information decreases, since $\boldsymbol{R}_x^{-1} = (\boldsymbol{R}_d + \boldsymbol{R}_v)^{-1}$. The numerical simulations in the following sections also reveal that

[4] $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite with $\boldsymbol{A}$ and $\boldsymbol{B}$ being Hermitian.

the unconditional CRB is always equal to or higher than the conditional CRB. Intuitively, when estimating the RTF, knowing the target signal itself would be more useful than knowing the signal statistics only. For further analytical insights, the reader can refer to [34].

VI. Experiments

In the preceding sections, we developed an RTF estimation algorithm that considers both spectral and spatial correlations. We computed conditional and unconditional CRBs to gauge achievable accuracy. Following this, we conduct simulations to compare the performance of our proposed wideband algorithm (SVD-direct) to the benchmark narrowband method (CW) and the established performance bounds. We employ two error metrics, the root-mean-squared-error (RMSE) and the Hermitian angle [24]. The RMSE is defined as:

$$\text{RMSE} = 10\log\sqrt{\frac{1}{KM}\|\hat{\boldsymbol{a}}-\boldsymbol{a}\|^2}\ (\text{dB}), \tag{28}$$

while the Hermitian angle is calculated as:

$$\frac{1}{K}\sum_{k=1}^{K}\text{acos}\left(\frac{|\,\hat{\boldsymbol{a}}_k^H\boldsymbol{a}_k|}{\|\hat{\boldsymbol{a}}_k^H\|\|\boldsymbol{a}_k\|}\right)\ (\text{rad}). \tag{29}$$

The RMSE accounts for discrepancies in the magnitude and phase, whereas the Hermitian angle depends exclusively on the angle between the RTFs. The CRBs are only defined for error measures based on the MSE. Therefore, these bounds are not shown in the plots that employ the Hermitian angle metric. We also define the signal-to-noise ratio (SNR) in the frequency domain as:

$$\text{SNR} = 10\log\frac{\sum_{i=1}^{KM}[\boldsymbol{R}_d]_{ii}}{\sum_{i=1}^{KM}[\boldsymbol{R}_v]_{ii}}\ (\text{dB}). \tag{30}$$

In all plots of Sections VI-A, VI-B, and VI-D, points connected by a continuous red line show the error for the proposed algorithm (Algorithm 1); points connected by a blue dotted line show errors for the benchmark algorithm (CW); points connected by a green dash-dotted line show the conditional CRB (Theorem 1); points connected by a purple dashed line show the unconditional CRB (Theorem 2).

We conduct five sets of experiments to explore increasingly realistic scenarios. In the first two sets of experiments (Sections VI-A and VI-B), we analyze scenarios where independent realizations of the signals are drawn from ideal multivariate Gaussian distributions. In Section VI-A, the target and noise powers at all frequencies are set to the same value and then rescaled to the desired SNR. Section VI-B describes a more realistic scenario where target and noise powers vary across frequencies. Results are shown for a single random draw of the target TF $\boldsymbol{a}$ and of the actual covariance matrices $\boldsymbol{R}_s$ and $\boldsymbol{R}_v$ because the CRB is defined for specific parameter values. Nonetheless, similar outcomes are observed for other realizations. To simulate the complex channel vector $\boldsymbol{a}$, we generate two uniformly distributed random vectors with values from -1 to 1 and use them for the real and imaginary parts. For the synthetic data of Sections VI-A and VI-B, the lines in the figures are the mean results averaged across 5000 Montecarlo realizations. The faded area represents the 95 % confidence interval [39]. The bounds are evaluated at the actual values of the parameters.

The other three sets of experiments deal with real data. The covariance matrices are thus estimated from overlapping STFT frames using the phase-corrected estimator introduced in Section III-A. Section VI-C investigates the correlation coefficients of measured speech signals. The experiments of Sections VI-D and VI-E apply the proposed algorithms to recorded anechoic speech convolved with real room impulse responses (RIRs), and evaluate both the RTF estimation accuracy and the effect of employing the estimated RTFs for beamforming. The ground truth TF $\boldsymbol{a}$ is computed as the discrete Fourier transform of the first K samples of the RIR. We perform 50 Montecarlo repetitions of the real-data experiments. Gaussian noise at 40 dB SNR is added to $\boldsymbol{v}$ in all experiments to account for sensor noise and simultaneously improve numerical conditioning of the inverse of the noise covariance matrix $\boldsymbol{R}_v$. We also measure the computational complexity of the algorithm in Section VI-F. As mentioned in Section I, all the simulations are implemented in Python, and the code to generate all figures in the paper is freely available online.

A. Equicorrelated, Equal Powers

The 'equicorrelated' formulation, also considered in [40], assumes that the noise signal exhibits identical variances at all sensors and frequency components. The target signal has unit variance at all frequency components. The cross-expectations over different frequency components are υ_f for the noise and ρ_f for the target. Because the frequency correlations are non-zero, the covariance matrices $\boldsymbol{R}_x$ and $\boldsymbol{R}_v$ describe non-WSS processes. Taking again the case of $M=2$ sensors and $K=2$ frequency components to simplify the exposition, the noise covariance matrix $\boldsymbol{R}_v$ is given by:

$$\boldsymbol{R}_v = \gamma^2\begin{bmatrix}1 & 0 & \upsilon_f & 0\\ 0 & 1 & 0 & \upsilon_f\\ \upsilon_f^* & 0 & 1 & 0\\ 0 & \upsilon_f^* & 0 & 1\end{bmatrix}, \tag{31}$$

where $\upsilon_f \in [0,1]$ and γ^2 is scaled according to (30) to yield the desired SNR. Similarly, the desired covariance matrix at the source $\boldsymbol{R}_s = \boldsymbol{R}_{\bar{s}} \otimes \boldsymbol{1}_{M\times M}$ is given by:

$$\boldsymbol{R}_s = \begin{bmatrix}1 & \rho_f\\ \rho_f^* & 1\end{bmatrix}\otimes\boldsymbol{1}_{M\times M} = \begin{bmatrix}1 & 1 & \rho_f & \rho_f\\ 1 & 1 & \rho_f & \rho_f\\ \rho_f^* & \rho_f^* & 1 & 1\\ \rho_f^* & \rho_f^* & 1 & 1\end{bmatrix}. \tag{32}$$

The desired covariance matrix at the receivers follows from (6). The stimuli $\boldsymbol{s}(l)$ and $\boldsymbol{v}(l)$, where l is the realization index, are generated through affine transformations applied to L independent and identically distributed realizations $\boldsymbol{n}(l)$ of a white complex multivariate Gaussian distribution $\boldsymbol{n}\sim\mathcal{CN}(\boldsymbol{0},\boldsymbol{I})$. For example, $\boldsymbol{s}(l) = \boldsymbol{R}_s^{1/2}\boldsymbol{n}(l)$, and this implies $\boldsymbol{s}(l)\sim\mathcal{CN}(\boldsymbol{0},\boldsymbol{R}_s)$. The estimates of the target and noise covariance matrices are derived through the sample covariance estimator of (9). The

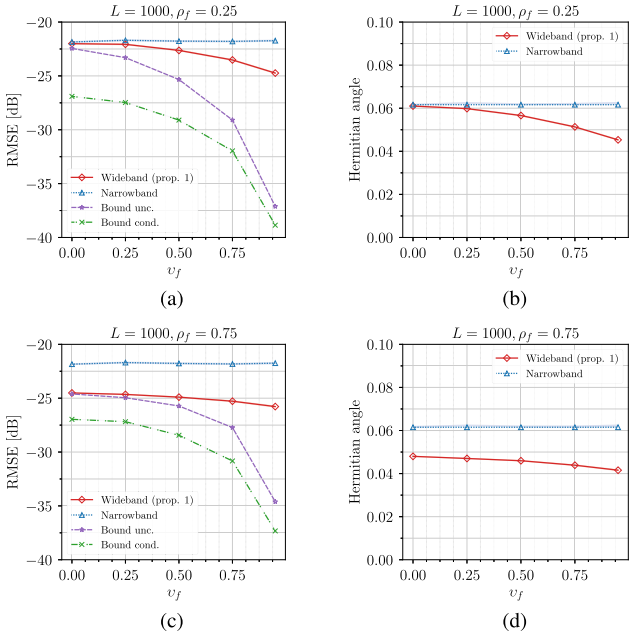

Fig. 3. Algorithm performance under varying noise frequency correlation υ_f, with different levels of target correlation ρ_f. The top two plots (a) and (b) represent a less correlated target ($\rho_f = 0.25$), while the bottom row (c) and (d) show a highly correlated target ($\rho_f = 0.75$). Each column corresponds to different evaluation metrics: the left column displays the RMSE, and the right column shows the Hermitian angle.

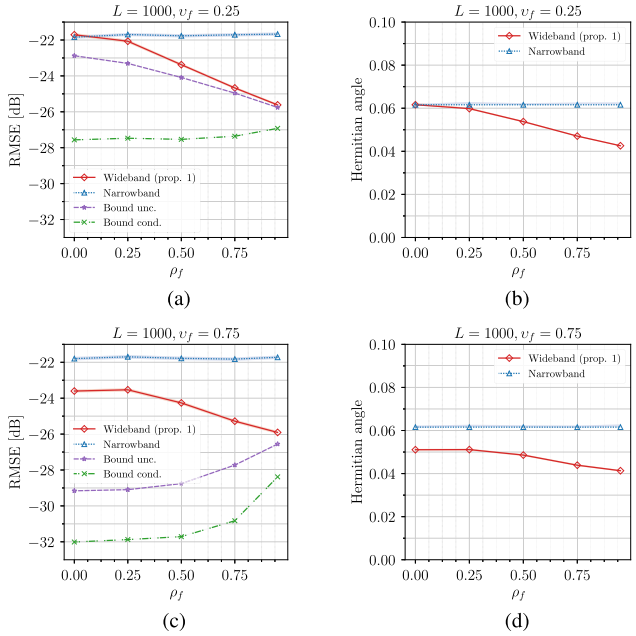

Fig. 4. Algorithm performance under varying target frequency correlation ρ_f, with different levels of noise correlation υ_f. The top two plots (a) and (b) represent less correlated noise ($\upsilon_f = 0.25$), while the bottom row (c) and (d) show highly correlated noise ($\upsilon_f = 0.75$). The left column corresponds to RMSE, and the right column shows the Hermitian angle.

phase-corrected estimator in (11) is indeed superfluous when independent signal realizations are available. Unless specified differently, the SNR is set to $-5\,\mathrm{dB}$ in all experiments. The signal correlation is set to $\rho_f = 0.25$, the noise correlation to $\upsilon_f = 0.25$, the number of frames to compute the sample covariance matrices to $L = 1000$, the number of sensors to $M = 2$ and the FFT length to $K = 5$. The true noise covariance matrix $\boldsymbol{R}_v$ is used in all algorithms, aligning with the CRB assumptions. Nonetheless, we noticed similar results when estimating $\boldsymbol{R}_v$ from a separate realization of the noise-only signal. The algorithms and the bounds are tested by varying four independent parameters: noise correlation υ_f, target correlation ρ_f, number of time frames L, and SNR.

1) Varying Noise Correlation υ_f: In the first experiment, we analyze the performance of the algorithms as the noise frequency correlation υ_f varies between 0 and 1 (Fig. 3). We generally observe that the RMSE and the Hermitian angle metrics follow similar trends. Let us first consider the scenario where the target has low correlation ($\rho_f = 0.25$), corresponding to Fig. 3(a) and (b). The two algorithms perform equally well when the noise correlation υ_f is low, while the proposed method shows improved accuracy for high values of υ_f. In other words, the SVD-direct algorithm can partially take advantage of increased noise correlation, while the benchmark algorithm cannot. Now, consider the case where the target shows high correlation ($\rho_f = 0.75$), corresponding to Fig. 3(c) and (d). The proposed method outperforms the benchmark for all values of υ_f, with improvements of approximately $3\,\mathrm{dB}$ in RMSE and $0.02\,\mathrm{rad}$ in Hermitian angle. Examining the conditional and unconditional CRBs, we note that substantial accuracy improvements are achievable when the noise exhibits a high correlation.

2) Varying Target Correlation ρ_f: In this section, we analyze the performance of the algorithms as the target frequency correlation ρ_f varies between 0 and 1 (Fig. 4). Because SVD-direct is explicitly designed to take advantage of spectral correlations in the target, we expect it to yield better performance for higher values of ρ_f. If the noise has low correlation ($\upsilon_f = 0.25$, corresponding to the top row in Fig. 4), the two algorithms perform equally well for low target correlation values ρ_f. Additionally, we observe that the proposed method can fully exploit the target correlation and shows improvements in the accuracy of up to $4\,\mathrm{dB}$ in RMSE and $0.02\,\mathrm{rad}$ in Hermitian angle for high values of ρ_f. The benchmark algorithm is not affected by variations in the target spectral correlation. Notice that the proposed algorithm achieves the CRB if a high target correlation is present, meaning that further improvements in accuracy in this scenario are not possible. Interestingly, the unconditional performance bound exhibits different trends for low and high noise correlation. The unconditional bound decreases with higher target correlations when the noise correlation is low (Fig. 4(a)). Conversely, the maximum accuracy is lower as the target correlation increases for high noise correlation (Fig. 4(c)). This aligns with findings from previous studies [40]. This seeming discrepancy can be better understood through analogy: when two point sources are located close together in space, they show maximal *spatial* correlation and exhibit similar correlation patterns, making it difficult to separate them. In our experiment, the noise and target sources have high *spectral* correlation, and they share the same correlation pattern (31) and (32). We might say that they are

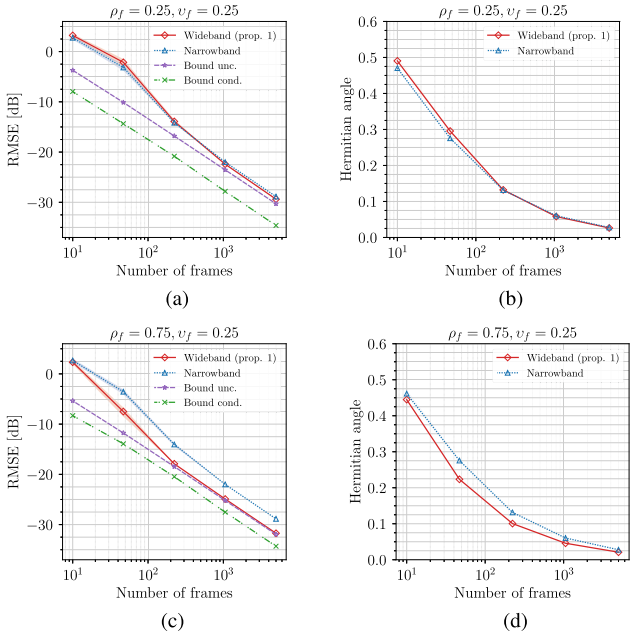

Fig. 5. Algorithm performance under varying number of time frames L, with different levels of target correlation ρ_f. The top two plots (a) and (b) represent a less correlated target ($\rho_f = 0.25$), while the bottom row (c) and (d) show a highly correlated target ($\rho_f = 0.75$). The left column corresponds to RMSE, and the right column shows the Hermitian angle.

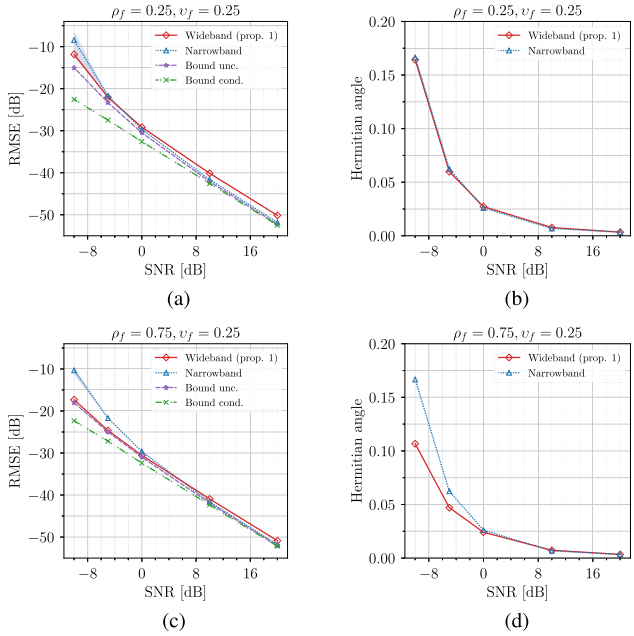

Fig. 6. Algorithm performance under varying SNR, with different levels of target correlation ρ_f. The top two plots (a) and (b) represent a less correlated target ($\rho_f = 0.25$), while the bottom row (c) and (d) show a highly correlated target ($\rho_f = 0.75$). The left column corresponds to RMSE, and the right column shows the Hermitian angle.

"spectrally superimposed" because their powers and correlation coefficients are the same, yielding very similar spectral covariance matrices. As a result, they are harder to distinguish than two spectrally independent sources.

3) Varying Number of Frames L: We now analyze the performance of the algorithms when the number of frames L to estimate the target covariance matrix $\boldsymbol{R}_d$ is varied between $L = 10$ and $L = 5000$ (Fig. 5). As expected, both algorithms perform better when more frames are available. For low values of target and noise correlation (Fig. 5(a) and (b)), the two algorithms perform similarly when the number of available time frames is large, whereas the proposed algorithm is slightly less accurate when L is small. When the target correlation is high, the wideband method shows smaller errors for any number of frames $L > 10$ and converges to the unconditional CRB for a high number of frames.

4) Varying SNR: This experiment analyzes the performance of the algorithms when the SNR varies between -10 and $20\,\mathrm{dB}$ (Fig. 6). Unsurprisingly, both algorithms perform better when the noise is less prominent. The methods perform similarly for low target correlation values (Fig. 6(a) and (b)). In contrast, for high target correlation (Fig. 6(c) and (d)), the proposed method shows significant performance gains of up $8\,\mathrm{dB}$ in RMSE and $0.05\,\mathrm{rad}$ in Hermitian angle in noisy scenarios. Both algorithms are close to the unconditional CRB for high SNR values.

B. Equicorrelated, Different Powers

In the second set of experiments, we extend the 'equicorrelated' scenario described in Section VI-A, by incorporating varying signal powers across different frequency components and sensors. This scenario is not only more realistic than the previous one, but it also leads to more diverse spectral correlation patterns — that is, covariance matrices — for the target and the noise signal, limiting the "spectral superposition" phenomenon observed in Section VI-A-2. In this simulation model, special care must be taken to ensure the validity of the simulated covariance matrices $\boldsymbol{R}_v$ and $\boldsymbol{R}_s$.

Let $[\boldsymbol{v}]_{km} = v_{km}$ be the noise signal at frequency k and sensor m, with variance $\mathbb{E}[|v_{km}|^2] = \gamma_{km}^2$. By the Cauchy–Schwarz inequality, it is known that the covariance between the two discrete complex random variables $v_{k_1 m_1}, v_{k_2 m_2}$, corresponding to the $(k_1 m_1, k_2 m_2)$ element of $\boldsymbol{R}_v$, is upper-bounded by:

$$|\,\mathbb{E}[v_{k_1 m_1} v_{k_2 m_2}^*]|^2 \leq \mathbb{E}[|v_{k_1 m_1}|^2]\,\mathbb{E}[|v_{k_2 m_2}|^2]. \tag{33}$$

Therefore, we can simulate $\boldsymbol{R}_v$ with a two-step procedure. First, the variances γ_{km}^2 on the diagonal of $\boldsymbol{R}_v$ are drawn from a uniform distribution $\mathcal{U}(\epsilon, 0.5)$, where $\epsilon > 0$ is a small positive number. Next, the covariance values are calculated as

$$\mathbb{E}[v_{k_1 m_1} v_{k_2 m_2}^*] = \begin{cases} 0, & \text{if } m_1 \neq m_2, \\ \upsilon_f \sqrt{\gamma_{k_1 m_1}^2 \gamma_{k_2 m_2}^2} & \text{if } m_1 = m_2, \end{cases} \tag{34}$$

where the factor $\upsilon_f \in [0, 1]$ models the noise inter-frequency correlation. Because $\upsilon_f \leq 1$, (34) leads to covariance values that are always smaller than their theoretical maxima. The correlations across different sensors are set to 0 since we model spatially uncorrelated noise. Finally, $\boldsymbol{R}_v$ is rescaled by a global noise variance γ^2 to yield the desired SNR according to (30). Analogously, the desired covariance matrix at the source is given

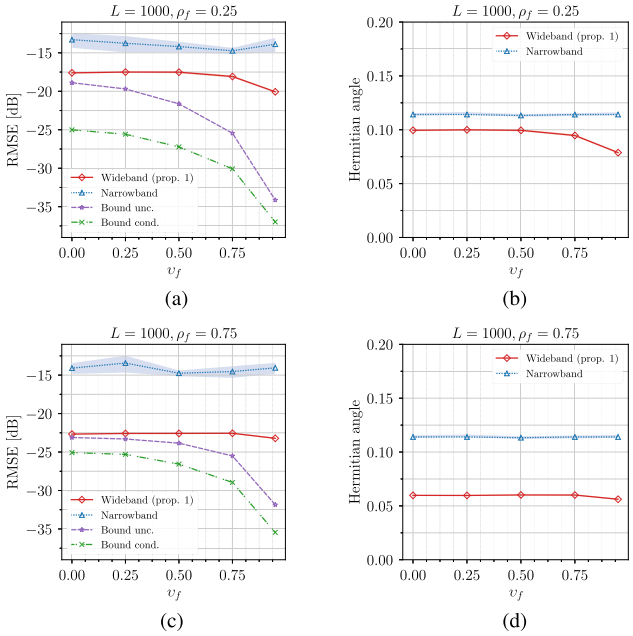

Fig. 7. Algorithm performance for non-uniform target and noise powers under varying noise frequency correlation υ_f, with different levels of target correlation ρ_f. The top two plots (a) and (b) represent a less correlated target ($\rho_f = 0.25$), while the bottom row (c) and (d) show a highly correlated target ($\rho_f = 0.75$). The left column corresponds to RMSE, and the right column shows the Hermitian angle.

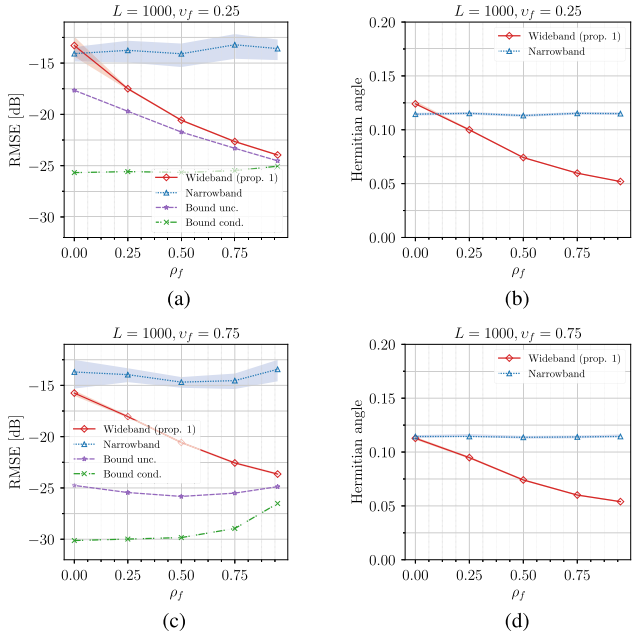

Fig. 8. Algorithm performance for non-uniform target and noise powers, under varying target frequency correlation ρ_f, with different levels of noise correlation υ_f. The top row ((a) and (b)) represents less correlated noise ($\upsilon_f = 0.25$), while the bottom row ((c) and (d)) shows highly correlated noise ($\upsilon_f = 0.75$). The left column corresponds to RMSE, and the right column shows the Hermitian angle.

by:

$$[\boldsymbol{R}_{\bar{s}}]_{k_1 k_2} = \begin{cases} \sigma^2_{k_1 k_2} \sim \mathcal{U}(\epsilon, 0.5) & \text{if } k_1 = k_2, \\ \rho_f \sqrt{\sigma^2_{k_1 k_1} \sigma^2_{k_2 k_2}} & \text{if } k_1 \neq k_2. \end{cases} \tag{35}$$

The desired covariance matrix at the receivers follows again from (6) and (7). The sampling procedure and the other simulation parameters follow from Section VI-A.

1) Varying Noise Correlation υ_f: The performance of the algorithms is examined as the noise frequency correlation υ_f varies from 0 to 1 (Fig. 7). The wideband algorithm outperforms the narrowband one in all cases. The difference in accuracy is larger when the target correlation ρ_f is higher (Fig. 7(c) and (d)), reaching an improvement of up to $8\,\mathrm{dB}$ RMSE and $0.05\,\mathrm{rad}$. Notice that the performance gains are more significant than in the 'equal powers' scenario of Section VI-A-1. We also observe that the error of the SVD-direct algorithm slightly decreases for very high noise correlation $\upsilon_f \geq 0.75$. Still, the gap between the unconditional CRB and the algorithms indicates that further improvements are possible.

2) Varying Target Correlation ρ_f: Next, we turn to one of the key experiments of the present study, where we analyze the performance of the algorithms as the target frequency correlation ρ_f varies between 0 and 1 for arbitrary noise and signal powers (Fig. 8). The wideband algorithm takes advantage of higher target spectral correlations ρ_f, as already observed in Section VI-A-2: both for low and high noise correlation, SVD-direct has significantly better performance for higher values of ρ_f, reaching improvements of $10\,\mathrm{dB}$ RMSE and $0.05\,\mathrm{rad}$ Hermitian angle. On the other hand, CW performs slightly better when the target correlation is completely absent ($\rho_f = 0$). A likely explanation is that the narrowband approach exploits the a priori knowledge that the target signal is uncorrelated across frequency. However, we argue that the scenario where $\rho_f = 0$ is unlikely to occur in practice for the reasons highlighted in Section I. This intuition is also confirmed in the correlation analysis of real data in Section VI-C. Turning our focus to the performance bounds, we notice that the SVD-direct method achieves the CRB if a high target correlation is present.

3) Varying Number of Frames L: We now evaluate the different approaches when estimating the covariance matrices with varying numbers of time frames L (Fig. 9). The narrowband and wideband approaches exhibit similar performance for scenarios with low target and noise correlation (Fig. 9(a) and (b)). The benchmark method is slightly more accurate when only a few frames are available ($L < 50$), whereas the proposed algorithm outperforms CW for a higher number of frames. Because wideband spectral-spatial covariance matrices are considerably larger than narrowband spatial covariance matrices, accurate estimation of the former requires more realizations. When the target is highly correlated (Fig. 9(c) and (d)), the wideband method consistently matches or outperforms the narrowband method.

4) Varying SNR: Lastly, in Fig. 10, we examine the performance for various SNR levels. For low target and noise spectral correlation (Fig. 10(a) and (b)), the two algorithms perform comparably, with the wideband method being marginally less accurate for higher SNRs. By contrast, when the target correlation is high, as in Fig. 10(c) and (d), the two approaches perform similarly for less noisy scenarios, but the wideband method has a

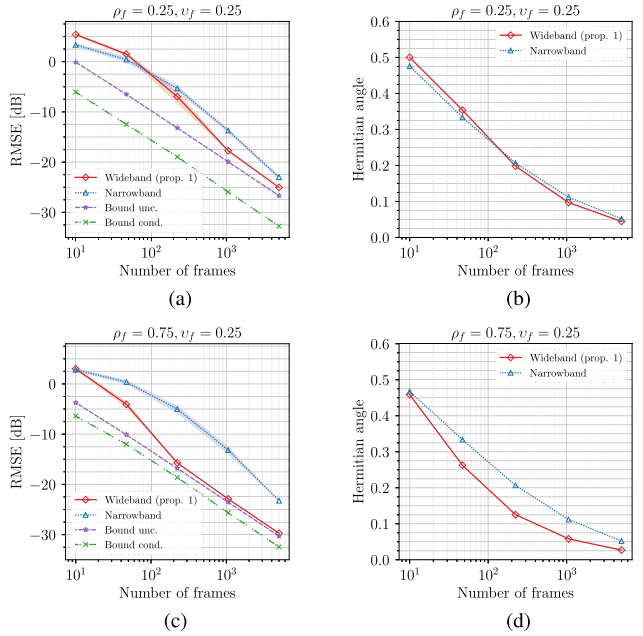

Fig. 9. Algorithm performance for non-uniform target and noise powers, under varying number of time frames L, with different levels of target correlation ρ_f. The top row ((a) and (b)) represents less correlated target ($\rho_f = 0.25$), while the bottom row ((c) and (d)) shows highly correlated target ($\rho_f = 0.75$). The left column corresponds to RMSE, and the right column shows the Hermitian angle.

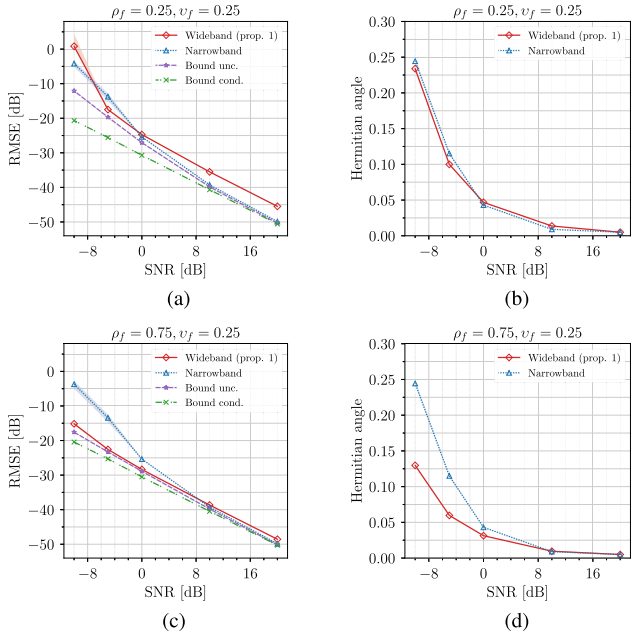

Fig. 10. Algorithm performance for non-uniform target and noise powers under varying SNR, with different levels of target correlation ρ_f. The top row ((a) and (b)) represents less correlated target ($\rho_f = 0.25$), while the bottom row ((c) and (d)) shows highly correlated target ($\rho_f = 0.75$). The left column corresponds to RMSE, and the right column shows the Hermitian angle.

considerably lower error for lower SNRs, with a reduction of up to 11 dB RMSE and 0.12 rad Hermitian angle at -10 dB SNR. This experiment concludes the evaluations on synthetic data.

C. Correlation Coefficients of Measured Data

Before testing the RTF estimation algorithms on real data, it is useful to examine the correlation coefficients of measured adult speech and white Gaussian noise, and relate them to the simulated coefficients from Sections VI-A and VI-B. To analyze the distribution of the spectral correlation coefficients, we first select a segment of length T from the time-domain signal of interest. After transforming this segment to the STFT domain, we estimate its spectral covariance matrix and spectral correlation coefficients, as detailed below.

The measured speech signal consists of anechoic male and female speech recordings from the Harvard Word List,[5] sampled at $f_s = 16$ kHz. The recordings last approximately 5 minutes. For each of the 50 Monte Carlo iterations, we randomly select a segment of length $T = 0.35$ s from either of the two recordings. Silent segments are discarded. The STFT analysis is performed with window length $K_2 = 1024$, corresponding to $K_+ = (K_2/2) + 1 = 513$ positive frequencies. We use a square-root Hann window function and a 75 % overlap between frames, such that the block-shift equals $R = 256$ samples. Therefore, the number of STFT frames available for estimating the spectral covariance matrices is $L \approx (T f_s - K_2 + R)/R \approx 20$. The number of frames L is thus small compared to the number of positive frequency bins K_+, complicating the estimation of the spectral covariance matrices. To focus the analysis on relevant frequency bands, we only consider frequency components between 0.08 kHz to 2.0 kHz, reducing the number of frequency bins from $K_+ = 513$ to $K \approx 124$. The remaining bins are ignored in the analysis.

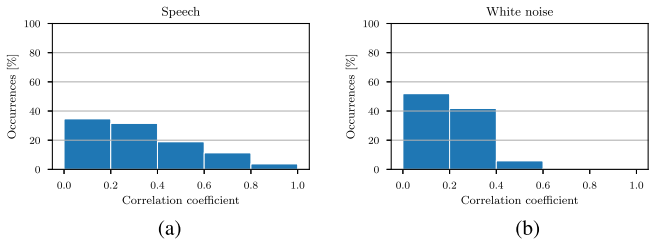

Fig. 11. Empirical distribution of spectral correlation coefficients for (a) male speech and (b) white noise signals. Speech has more highly correlated bins than white noise, a stationary signal.

The spectral covariance matrix $\hat{\boldsymbol{R}} \in \mathbb{C}^{K \times K}$ is estimated from phase-adjusted STFT data, as described in (11), to account for the phase shifts caused by overlapping frames. The magnitudes of the correlation coefficients for the off-diagonal elements are then obtained as:

$$\hat{\rho}_{f,k_1k_2} = \left| [\hat{\boldsymbol{R}}]_{k_1k_2} / \sqrt{\hat{\sigma}^2_{k_1k_1} \hat{\sigma}^2_{k_2k_2}} \right|, \tag{36}$$

where $k_1, k_2 = 1, \ldots, K$, $k_1 \neq k_2$, and $\hat{\sigma}^2_{k_1k_1} = [\hat{\boldsymbol{R}}]_{k_1k_1}$. Finally, we count the number of bins within each interval of the histogram and average the percentages across Monte Carlo realizations (Fig. 11). As expected, the speech data (Fig. 11(a)) exhibits significantly higher spectral correlation than the white

[5] "Speech Intelligibility CD" from Neil Thompson Shade.

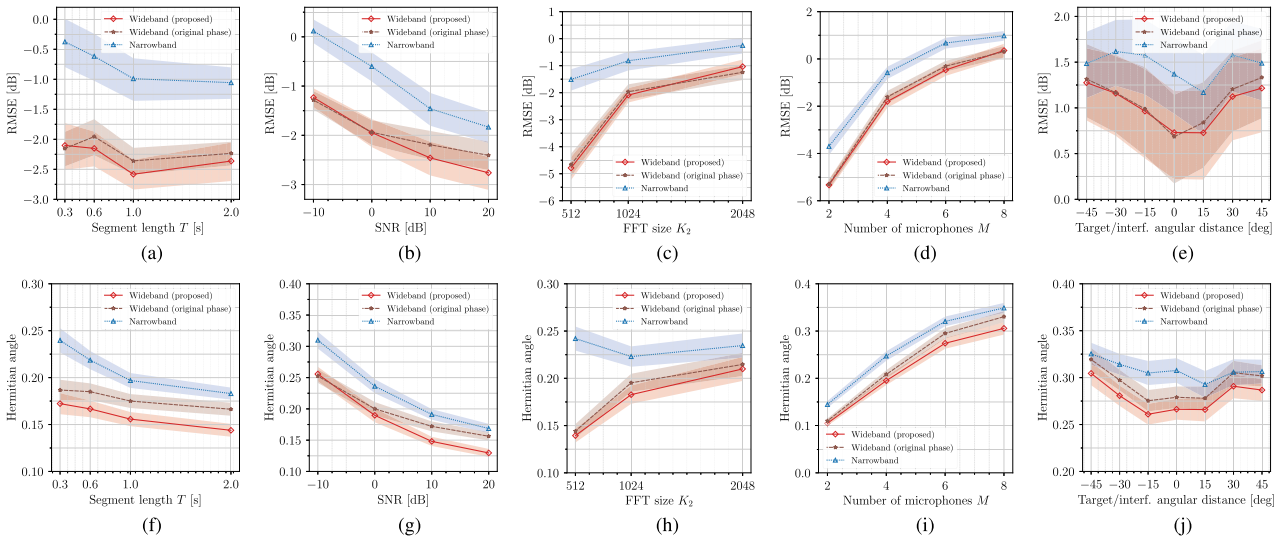

Fig. 12. Performance of the algorithms for real speech, under (a-f) varying segment length T, (b-g) varying SNR, (c-h) varying FFT size, (d-i) varying number of microphones, and (e-j) varying angular distance between target and interferer. The error metrics are the RMSE (top) and the Hermitian angle (bottom).

noise data (Fig. 11(b)). Approximately 20% of the speech data shows correlations above 0.6, while this value is nearly zero for the white noise data. Surprisingly, the white noise data displays spectral correlations exceeding 0.2 in over 40% of the cases, which we hypothesize is due to spectral leakage caused by the finite-length windowing effect. Given that the proposed RTF estimation method performs better for highly correlated target signals, the large number of low-correlation bins suggests that RTF estimation may improve by applying the proposed algorithm to the highly correlated bins only. The optimal design of such a method should be explored in future research.

D. Real-Data Simulations

The fourth set of experiments tests the RTF estimation algorithms on real speech data, maintaining the same settings as in Section VI-C, except where noted. A directional interferer is introduced by randomly sampling a real-world recording from the ESC-50 database [41]. Within the ESC-50 database, we sample from three selected categories that contain approximately stationary sounds: engine noise, washing machine noise, and vacuum cleaner noise. The default SNR for the interferer is set to $0\,\mathrm{dB}$. The target and the interfering signals are generated by convolution with the RIRs from the database in [42]. The RIRs were measured with a linear microphone array with 8 sensors, spaced $8\,\mathrm{cm}$ apart, in a room of size of $6 \times 6 \times 2.4\,\mathrm{m}$. The average reverberation time of the room is $RT_{60} = 0.61\,\mathrm{s}$. All RIRs are cut after $0.61\,\mathrm{s}$ to reduce the length of the convolved signals while preserving most of the reverberation power. Unless otherwise specified, only the first $M = 4$ microphones of the array are used. The target and interfering sources are placed $1\,\mathrm{m}$ away from the microphones, at angles of $45\,^\circ$ and $60\,^\circ$, respectively. Therefore, the target and interfering sources are spatially close to each other but exhibit different spectral properties. The noise covariance matrix $\hat{\boldsymbol{R}}_v$ is estimated from a separate realization of the noise-only signal whose total duration is $T_n = 2\,\mathrm{s}$. A distinct noise realization is used for each Monte Carlo iteration. The covariance matrices $\hat{\boldsymbol{R}}_x$ and $\hat{\boldsymbol{R}}_v$ are estimated from phase-adjusted STFT data, as described in (11), to account for the phase shifts caused by overlapping frames. To gauge the improvements brought by the phase-corrected covariance estimator, we also depict the accuracy of the wideband SVD-direct algorithm when utilizing the sample covariance estimator of (9). Errors based on the sample covariance estimate with the original phase values are indicated by appending "original phase" to the RTF estimator name. When calculating the errors on the RTF estimates, we only retain the frequency bands for which the average power of the target signal at the microphones is no more than $35\,\mathrm{dB}$ lower than that of the loudest frequency band. Notice that the experiments include all bands in the specified frequency and SNR region. That means even those bands without correlation across frequency are included.

The wideband and narrowband algorithms are assessed based on the RMSE and the Hermitian angle metrics across different conditions. We analyze the algorithms for different variations in the segment length T (Fig. 12(a) and (f)), which determines the number of frames L available for estimating the covariance matrices, the SNR (Fig. 12(b) and (g)), the FFT size K_2 (Fig. 12(c) and (h)), hence K, the number of microphones M (Fig. 12(d) and (i)), and the directional interferer angular position (Fig. 12(e) and (j)). The proposed phase-adjusted wideband algorithm outperforms the narrowband benchmark in all experiments of Fig. 12. The performance gap between wideband and narrowband algorithms remains largely unchanged under varying conditions. Notice that phase correction would not affect the performance of the narrowband algorithm, CW, which does not rely on inter-frequency correlations. On the other hand, the wideband algorithm SVD-direct benefits significantly from incorporating phase-adjusted covariance matrices, especially in

scenarios with a higher number of available time frames L or microphones M, or when the SNR is high. The impact of phase correction appears to diminish under conditions where the covariance estimates are compromised due to a low number of time frames or a low SNR. In Fig. 12(d) and (i), we observe a decline in the performance of all algorithms as the number of microphones increases. This is likely because, as M increases linearly, the number of elements in the covariance matrices ($M \times M$) increases quadratically. Consequently, with the data length remaining constant, the quality of the estimated covariance matrices worsens. Fig. 12(e) and (j) illustrate the RTF estimation errors when the target is positioned at 45° and the interferer is placed at various angles within the 0° to 90° range. Although the wideband method outperforms the narrowband approach, the performance gap is greatest when the target and interferer are separated by a narrow angle, diminishing as the angular distance increases. Exploiting spectral correlations proves to be most effective when there is significant overlap in the spatial correlations of the target and interferer.

E. Beamforming

Knowledge of the RTF of a target speaker allows us to virtually steer a beamformer towards them and enhance the quality and the intelligibility of speech. In this section, we evaluate the performance of an MVDR beamformer that uses the estimated RTFs to enhance a target signal, to get an impression of the performance improvement by using the proposed RTF estimator. The output of the beamformer is an estimate of the target signal and its early reflections at the reference microphone, given by:

$$\hat{d}_{1,k}(l) = \boldsymbol{w}_k^H \boldsymbol{x}_k(l), \quad l = 1, \ldots, L \text{ and } k = 1, \ldots, K, \quad (37)$$

where $\boldsymbol{w}_k \in \mathbb{C}^M$ are the beamforming weights, given by

$$\boldsymbol{w}_k = \frac{\hat{\boldsymbol{R}}_n^{-1}(k)\hat{\boldsymbol{a}}_k}{\hat{\boldsymbol{a}}_k^H \hat{\boldsymbol{R}}_n^{-1}(k)\hat{\boldsymbol{a}}_k}, \quad k = 1, \ldots, K. \quad (38)$$

In Fig. 13, we compare the output of the MVDR beamformer at various SNRs, using four different RTF estimates $\hat{\boldsymbol{a}}_k$: SVD-direct (wideband), CW (narrowband), the true RTF $\boldsymbol{a}_k$, and the unprocessed one at the reference microphone, i.e., $\hat{\boldsymbol{a}}_k = [1, 0, \ldots, 0]^T$. The output $\hat{d}_{1,k}(l)$ is evaluated using the short-time intelligibility index (STOI), the log-likelihood-ratio (llr) spectral distance, and the frequency-weighted segmental SNR (fwSNRseg) [43], [44]. The same settings as in Sections VI-C and VI-D are used, except for the segment length T, which is extended to $T = 1$ s. As expected, the MVDR beamformer performs best when using the true RTF, while using the noisy signal at the reference microphone results in the worst scores across all metrics. The proposed algorithm generally outperforms CW according to all metrics in most conditions, except at very low SNRs.

F. Computational Complexity

Let us now compare the computational complexity of SVD-direct with the benchmark algorithm, CW. The cost of estimating

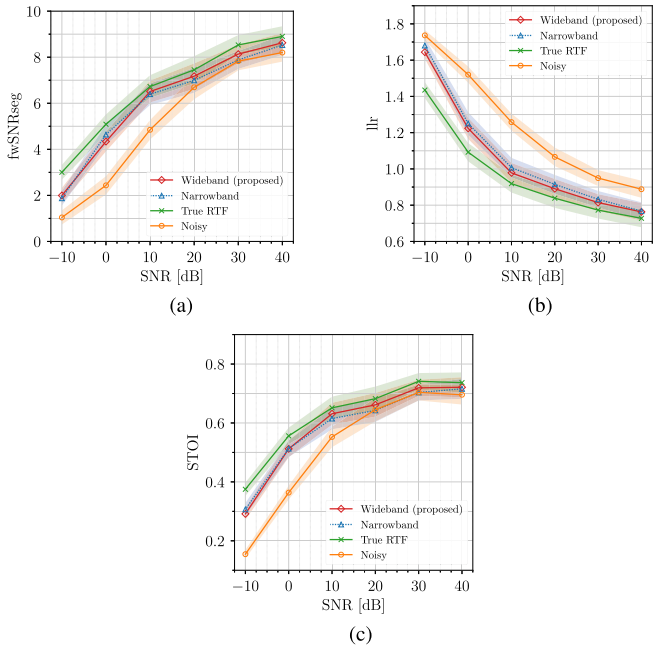

Fig. 13. Evaluation of MVDR beamformer outputs using different RTF estimates, under varying SNR. Different subfigures correspond to different metrics: fwSNRseg (a), LLR (b), and STOI (c).

the noisy covariance matrix is neglected as it is not considered part of the algorithms. While CW applies K generalized eigendecompositions on spatial $M \times M$ covariance matrices, SVD-direct requires an initial generalized eigendecomposition on two matrices of size $KM \times KM$, followed by K SVDs on matrices of size $M \times KM$. Consequently, SVD-direct is slower than CW, primarily because the computational cost of eigenvalue decomposition increases cubically with the matrix size. Keeping the same settings as in the real-data experiments of Section VI-D, our measurements reveal that CW is approximately 200 times faster than SVD-direct using our non-optimized Python implementation on a MacBook Pro 16 inches (M1 Max chip).

VII. Additional Discussion

Our experiments yielded valuable insights into the performance of the narrowband CW and the wideband SVD-direct methods for RTF estimation, comparing them with the conditional and the unconditional performance bounds. Let us now examine the key findings.

Our investigation reveals a consistent trend favoring the wideband approach in scenarios with higher target correlation. Unlike the narrowband method, which remains unaffected by varying noise spectral correlation, the wideband approach also demonstrates occasional performance improvements when dealing with highly correlated noises.

The CRB analysis reveals a fascinating insight: when the noise spectral correlation υ_f is high, significant potential exists for further improvements in wideband channel estimation. While this discovery pertains to channel estimation, it points to potential applications across various parameter estimation tasks. The finding also provides a theoretical foundation for the observed

empirical evidence in certain parametric and machine-learning approaches. Notably, some DNNs that operate on the entire time-frequency representation outperform narrowband alternatives in various speech enhancement tasks [25], [28]. It also offers a partial explanation for the intelligibility gains experienced by humans when detecting speech affected by harmonic noises [16].

The correlation analysis of Section VI-C, together with the real-speech experiments in Sections VI-D and VI-E, confirm that natural speech possesses spectral correlations that can be exploited by the SVD-direct algorithm, leading to improved RTF estimation and beamforming performance in the scenarios under analysis. However, it is worth noting that the wideband algorithm may not surpass narrowband algorithms in certain settings, such as when dealing with highly non-stationary noise sources. This limitation arises from the inherent challenge of estimating large spectral-spatial covariance matrices from a limited number of frames. For instance, typical speech has about 10-20 different sounds per second [45, Chapter 15.3]. This dynamic nature makes the estimation of spectral patterns more demanding compared to spatial patterns, which depend on the positions of the speaker and the listener. Improved spectral correlation estimates may result by modeling the signals under analysis as realizations of cyclostationary processes, a particular class of spectrally correlated processes [46].

VIII. Conclusion

The uncorrelation of frequency components of a signal is a ubiquitous assumption that is often not verified in practice due to STFT processing and the non-stationary nature of signals. In this paper, we investigated the role of spectral correlations in spatial processing and proposed a new subspace-based algorithm for the channel estimation task. Indeed, accurate knowledge of the acoustic transfer functions between target speakers and microphones is crucial for spatial filtering in applications like MVDR beamforming.

Extensive numerical experiments demonstrated the superior performance of our wideband approach over the maximum-likelihood narrowband benchmark, yielding gains of more than $10\,\mathrm{dB}$ RMSE in scenarios involving spectral correlations and low SNR. The proposed SVD-direct algorithm also exhibited competitive performance with real reverberant speech data contaminated by directional interferers and spatially uncorrelated noise. These achievements are obtained without compromising conceptual interpretability, as the channel estimate can be computed in closed form with just a few lines of code.

Furthermore, we derived CRBs for wideband channel estimation, revealing the potential for substantial accuracy improvements when noise, and to a lesser extent, the target exhibits high-frequency correlation. This study serves as a starting point for understanding the impact of spectral correlations on parameter estimation for array processing. Future endeavors will focus on refining the estimation of spectral-spatial covariance matrices, conducting more comprehensive tests on real-world measurements, and analyzing the influence of spectral correlation in speech enhancement and acoustic source separation.

Appendix A
Proof of Conditional CRB [Equation (25)]

Proof: To obtain the Cramér–Rao bound for the unknown parameters $\boldsymbol{\theta}$, we first calculate the derivatives of the log-likelihood function with respect to the unknown parameters to form the Fisher information matrix. Notice that the log-likelihood is a real-valued function of a complex variable $\boldsymbol{\theta}$. Thus, by evaluating the gradients using Wirtinger derivatives [37], we can make use of the following properties.

Lemma 2: Let $f : \mathbb{C}^p \times \mathbb{C}^p \times \mathbb{C}^q \times \mathbb{C}^q \mapsto \mathbb{R}$ be a real scalar function of four complex variables $\boldsymbol{w}, \boldsymbol{w}^* \in \mathbb{C}^p$ and $\boldsymbol{z}, \boldsymbol{z}^* \in \mathbb{C}^q$. Then

a) $\nabla_{\boldsymbol{z}} f = (\nabla_{\boldsymbol{z}^*} f)^*$.
b) $\nabla_{\boldsymbol{z}} \nabla_{\boldsymbol{w}}^H f = (\nabla_{\boldsymbol{z}^*} \nabla_{\boldsymbol{w}^*}^H f)^*$.

Proof:

a) Follows from the fact the gradient operators are complex conjugates while f is real.
b)

$$\begin{aligned}\nabla_{\boldsymbol{z}} \nabla_{\boldsymbol{w}}^H f &= \nabla_{\boldsymbol{z}} \nabla_{\boldsymbol{w}^*}^T f = (\nabla_{\boldsymbol{w}^*} \nabla_{\boldsymbol{z}}^T)^T f \\ &= (\nabla_{\boldsymbol{w}^*} \nabla_{\boldsymbol{z}^*}^H)^T f = (\nabla_{\boldsymbol{z}^*} \nabla_{\boldsymbol{w}^*}^H f)^*.\end{aligned}$$

□

For notational convenience, we define $\mathcal{L}(\boldsymbol{\theta}) = \ln p(\boldsymbol{X}; \boldsymbol{\theta})$. We will begin by evaluating the bottom-right quadrant of the Fisher information matrix (20), defined as $-\mathbb{E}[\nabla_{\boldsymbol{a}^*} \nabla_{\boldsymbol{a}^*}^H \mathcal{L}(\boldsymbol{\theta})]$. Expanding the matrix product, the partial derivative of the log-likelihood $\mathcal{L}(\boldsymbol{\theta})$ in (24) with respect to a_k^* is given by

$$\nabla_{a_k^*} \mathcal{L}(\boldsymbol{\theta}) = -\nabla_{a_k^*} \left(\sum_{l=1}^{L} (\boldsymbol{x}(l) - \boldsymbol{A}\boldsymbol{s}(l))^H \boldsymbol{R}_v^{-1} \boldsymbol{v}(l) \right) \tag{39}$$

$$= \nabla_{a_k^*} \left(\sum_{l=1}^{L} \sum_{j=1}^{KM} s_j^*(l) \boldsymbol{a}_j^H \boldsymbol{R}_v^{-1} \boldsymbol{v}(l) \right) \tag{40}$$

$$= \sum_{l=1}^{L} s_k^*(l) \boldsymbol{e}_k^T \boldsymbol{R}_v^{-1} \boldsymbol{v}(l). \tag{41}$$

The second order derivative evaluates to

$$\nabla_{a_k^*} \nabla_{a_m} \mathcal{L}(\boldsymbol{\theta}) = -\sum_{l=1}^{L} s_k^*(l) \boldsymbol{e}_k^T \boldsymbol{R}_v^{-1} \boldsymbol{e}_m s_m(l). \tag{42}$$

This leads to

$$-\mathbb{E}[\nabla_{\boldsymbol{a}^*} \nabla_{\boldsymbol{a}^*}^H \mathcal{L}(\boldsymbol{\theta})] = \sum_{l=1}^{L} \boldsymbol{S}(l)^H \boldsymbol{R}_v^{-1} \boldsymbol{S}(l), \tag{43}$$

where we defined $\boldsymbol{S}(l) = \mathrm{diag}(\boldsymbol{s}(l))$. With this, the Fisher information matrix is given by (cf. Lemma 2) $\boldsymbol{I}_{\boldsymbol{\theta}} = \mathrm{blkdiag}(\boldsymbol{B}^*, \boldsymbol{B})$, where $\mathrm{blkdiag}(\cdot)$ is the operator that constructs a block diagonal matrix from the given matrices, and $\boldsymbol{B} = \sum_{l=1}^{L} \boldsymbol{S}(l)^H \boldsymbol{R}_v^{-1} \boldsymbol{S}(l)$. The block-diagonal matrix $\boldsymbol{I}_{\boldsymbol{\theta}}$ can be inverted block-wise, leading to

$$\boldsymbol{I}_{\boldsymbol{\theta}}^{-1} = \mathrm{blkdiag}((\boldsymbol{B}^*)^{-1}, \boldsymbol{B}^{-1}). \tag{44}$$

The variance of unbiased estimators of the ATF is, therefore, bounded by $\mathrm{var}(\hat{a}_i) \geq [(B^*)^{-1}]_{ii}, \quad i = 1, \ldots, M.$

Transfer functions can be estimated in relation to a reference sensor r with a function $\boldsymbol{g}(\cdot)$ defined in (22). Choosing $r = 1$ as a reference sensor, the Jacobian matrix can be written as

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{g} = \begin{bmatrix}\nabla_{\boldsymbol{a}}\boldsymbol{g} & \nabla_{\boldsymbol{a}^*}\boldsymbol{g}\end{bmatrix} = \begin{bmatrix}\nabla_{\boldsymbol{a}}\boldsymbol{g} & \boldsymbol{0}_{KM\times KM}\end{bmatrix}, \tag{45}$$

where the right block of the gradient, $\nabla_{\boldsymbol{a}^*}\boldsymbol{g}$, is null because $\boldsymbol{g}(\cdot)$ does not depend on $\boldsymbol{a}^*$. We can further partition the left block of the gradient in K "fat" matrices $\nabla_{\boldsymbol{a}}\boldsymbol{g} = [\nabla^T_{\boldsymbol{a}_1}\boldsymbol{g}, \ldots, \nabla^T_{\boldsymbol{a}_K}\boldsymbol{g}]^T$, where $\nabla_{\boldsymbol{a}_k}\boldsymbol{g} \in \mathbb{C}^{M\times KM}$, $k = 1, \ldots, K$. Individual blocks $\nabla_{\boldsymbol{a}_k}\boldsymbol{g}$ can be written as (46) shown at the bottom of this page so that $\nabla_{\boldsymbol{a}}\boldsymbol{g}$ shows a block-diagonal structure. With this, we have from (44) and (21)

$$\boldsymbol{R}_{\hat{\phi}} \succeq (\nabla_{\boldsymbol{\theta}}\boldsymbol{g})\boldsymbol{I}^{-1}_{\boldsymbol{\theta}}(\nabla^H_{\boldsymbol{\theta}}\boldsymbol{g}) = (\nabla_{\boldsymbol{a}}\boldsymbol{g})(\boldsymbol{B}^*)^{-1}(\nabla^H_{\boldsymbol{a}}\boldsymbol{g}). \tag{47}$$

The CRB corresponds to the diagonal elements of the matrix at the right-hand side of (47), as stated in (25). □

Appendix B
Proof of Unconditional CRB [Equation (27)]

We first list some derivative rules (e.g., [35]) for a generic square matrix $\boldsymbol{X}(\boldsymbol{\theta})$, where the values of $\boldsymbol{X}$ depend on $\boldsymbol{\theta}$.

$$\nabla_{\theta_i}\ln(\boldsymbol{X}) = \operatorname{tr}(\boldsymbol{X}^{-1}\nabla_{\theta_i}\boldsymbol{X}), \tag{48}$$

$$\nabla_{\theta_i}\operatorname{tr}(\boldsymbol{X}) = \operatorname{tr}(\nabla_{\theta_i}\boldsymbol{X}), \tag{49}$$

$$\nabla_{\theta_i}\boldsymbol{X}^{-1} = -\boldsymbol{X}^{-1}(\nabla_{\theta_i}\boldsymbol{X})\boldsymbol{X}^{-1}. \tag{50}$$

The computation of the unconditional CRB requires the calculation of first- and second-order derivatives of the log-likelihood in (26), reproduced here for easy reference:

$$\mathcal{L}(\boldsymbol{\theta}) = -L\ln|\pi\boldsymbol{R}_x| - L\operatorname{tr}(\hat{\boldsymbol{R}}_x\boldsymbol{R}_x^{-1}). \tag{51}$$

Proof: The partial derivative of the log-likelihood $\mathcal{L}(\boldsymbol{\theta})$ in (26) with respect to a^*_k is given by

$$\nabla_{a^*_k}\mathcal{L}(\boldsymbol{\theta}) = -L\operatorname{tr}(\boldsymbol{R}_x^{-1}\boldsymbol{F}_k) - L\operatorname{tr}(\boldsymbol{R}_x^{-1}\boldsymbol{F}_k\boldsymbol{R}_x^{-1}\hat{\boldsymbol{R}}_x), \tag{52}$$

where we introduced

$$\boldsymbol{F}_k = \nabla_{a^*_k}\boldsymbol{R}_x = \boldsymbol{A}\boldsymbol{R}_s\boldsymbol{E}^{kk} \tag{53}$$

and $\boldsymbol{E}^{ij}$ is zero everywhere and 1 at entry ij. The first term on the right-hand side of (52) is obtained directly from (48). The second term is obtained by using the derivative of the trace and of the matrix inverse as given by (49) and (50), respectively, along with the cyclic property of the trace operator. It is important to note that the estimate $\hat{\boldsymbol{R}}_x$ is independent of the parameter vector $\boldsymbol{a}$. The second-order partial derivative of $\mathcal{L}(\boldsymbol{\theta})$ writes

$$\nabla_{a_m}\nabla_{a^*_k}\mathcal{L}(\boldsymbol{\theta}) = -L\nabla_{a_m}[\operatorname{tr}(\boldsymbol{R}_x^{-1}\boldsymbol{F}_k) + \operatorname{tr}(\boldsymbol{R}_x^{-1}\boldsymbol{F}_k\boldsymbol{R}_x^{-1}\hat{\boldsymbol{R}}_x)]. \tag{54}$$

The derivatives of the two terms can be evaluated separately. For the first term in (54), we have

$$\begin{aligned}\nabla_{a_m}\operatorname{tr}(\boldsymbol{R}_x^{-1}\boldsymbol{F}_k) &= \operatorname{tr}(\nabla_{a_m}\boldsymbol{R}_x^{-1}\boldsymbol{F}_k)\\ &= \operatorname{tr}(-\boldsymbol{R}_x^{-1}\boldsymbol{G}_m\boldsymbol{R}_x^{-1}\boldsymbol{F}_k + \boldsymbol{R}_x^{-1}\boldsymbol{H}_{mk}),\end{aligned} \tag{55}$$

which is obtained by applying the product rule together with (49) and (50). Here, $\boldsymbol{G}_m$ and $\boldsymbol{H}_{mk}$ are defined as

$$\boldsymbol{G}_m = \nabla_{a_m}\boldsymbol{R}_x = \boldsymbol{E}^{mm}\boldsymbol{R}_s\boldsymbol{A}^H, \tag{56}$$

$$\boldsymbol{H}_{mk} = \nabla_{a_m}\boldsymbol{F}_k = \nabla_{a_m}\nabla_{a^*_k}\boldsymbol{R}_x = \boldsymbol{E}^{mm}\boldsymbol{R}_s\boldsymbol{E}^{kk}. \tag{57}$$

The second term in (54) is given by

$$\begin{aligned}&\nabla_{a_m}\operatorname{tr}(\boldsymbol{R}_x^{-1}\boldsymbol{F}_k\boldsymbol{R}_x^{-1}\hat{\boldsymbol{R}}_x) = \operatorname{tr}\nabla_{a_m}(\boldsymbol{R}_x^{-1}\hat{\boldsymbol{R}}_x\boldsymbol{R}_x^{-1}\boldsymbol{F}_k)\\ &= \operatorname{tr}\left[\boldsymbol{R}_x^{-1}\hat{\boldsymbol{R}}_x\boldsymbol{R}_x^{-1}(\boldsymbol{G}_m\boldsymbol{R}_x^{-1}\boldsymbol{F}_k - \boldsymbol{H}_{mk} + \boldsymbol{F}_k\boldsymbol{R}_x^{-1}\boldsymbol{G}_m)\right],\end{aligned} \tag{58}$$

which was again obtained by utilizing the product rule, (49), (50), and rearranging the resulting terms. By combining (55) and (58), the negative expected second-order partial derivative follows as

$$\begin{aligned}&-\mathbb{E}[\nabla_{a_m}\nabla_{a^*_k}\mathcal{L}(\boldsymbol{\theta})]\\ &= L\,\mathbb{E}\left\{\operatorname{tr}\left[-\boldsymbol{R}_x^{-1}(\boldsymbol{G}_m\boldsymbol{R}_x^{-1}\boldsymbol{F}_k - \boldsymbol{H}_{mk})\right]\right.\\ &\left.+\operatorname{tr}\left[\boldsymbol{R}_x^{-1}\hat{\boldsymbol{R}}_x\boldsymbol{R}_x^{-1}(\boldsymbol{G}_m\boldsymbol{R}_x^{-1}\boldsymbol{F}_k - \boldsymbol{H}_{mk} + \boldsymbol{F}_k\boldsymbol{R}_x^{-1}\boldsymbol{G}_m)\right]\right\}\\ &= L\operatorname{tr}\left(\boldsymbol{R}_x^{-1}\boldsymbol{F}_k\boldsymbol{R}_x^{-1}\boldsymbol{G}_m\right).\end{aligned} \tag{59}$$

To collect the expected values of the second-order partial derivative, we define a matrix $\boldsymbol{C}_1$ such that $[\boldsymbol{C}_1]_{mk} = -\mathbb{E}[\nabla_{a_m}\nabla_{a^*_k}\mathcal{L}(\boldsymbol{\theta})]$. The elements of the bottom left block of the Fisher information matrix can be similarly obtained as $[\boldsymbol{C}_2]_{mk} = -\mathbb{E}[\nabla_{a_m}\nabla_{a_k}\mathcal{L}(\boldsymbol{\theta})] = L\operatorname{tr}(\boldsymbol{R}_x^{-1}\boldsymbol{G}_k\boldsymbol{R}_x^{-1}\boldsymbol{G}_m)$, and the elements $-\mathbb{E}[\nabla_{a^*_m}\nabla_{a^*_k}\mathcal{L}(\boldsymbol{\theta})]$ of the top right block follow as $\boldsymbol{C}_2^H$ from Lemma 2. The inverse of the Fisher information matrix for the unconditional case can then be represented by:

$$\boldsymbol{I}^{-1}_{\boldsymbol{\theta}} = \begin{bmatrix}\boldsymbol{C}_1^* & \boldsymbol{C}_2^H\\ \boldsymbol{C}_2 & \boldsymbol{C}_1\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{C} & *\\ * & *\end{bmatrix}, \tag{60}$$

where $\boldsymbol{C} \in \mathbb{C}^{KM\times KM}$ is obtained by selecting the first KM rows and columns from $\boldsymbol{I}^{-1}_{\boldsymbol{\theta}}$. To derive the bound for estimating the *relative* transfer function, we employ the mapping $\boldsymbol{g}(\cdot)$ as defined in (22). Using (21), the inverse Fisher information matrix is left- and right-multiplied by $\nabla_{\boldsymbol{\theta}}\boldsymbol{g}$, which is defined in (45), resulting in the final form of the bound given by (27). □

$$\left[\begin{array}{c|ccccc|c}\multirow{5}{*}{$\boldsymbol{0}_{M\times(k-1)M}$} & 0 & 0 & 0 & & 0 & \multirow{5}{*}{$\boldsymbol{0}_{M\times(K-k)M}$}\\ & -a_{k2}a_{k1}^{-2} & a_{k1}^{-1} & 0 & \ldots & 0 & \\ & -a_{k3}a_{k1}^{-2} & 0 & a_{k1}^{-1} & \ldots & 0 & \\ & \vdots & & & \ddots & \vdots & \\ & -a_{kM}a_{k1}^{-2} & 0 & \ldots & 0 & a_{k1}^{-1} & \end{array}\right], \tag{46}$$

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic Beamforming for Hearing Aid Applications," in *Handbook on Array Processing and Sensor Networks*. Hoboken, NJ, USA: Wiley, 2010, pp. 269–302.

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

[6] 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.

[7] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint maximum likelihood estimation of power spectral densities and relative acoustic transfer functions for acoustic beamforming," in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech Signal Process.*, Jun. 2021, pp. 6119–6123.

[8] C. Li, J. Martinez, and R. C. Hendriks, "Low complex accurate multi-source RTF estimation," in *Proc. ICASSP 2022-2022 IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2022, pp. 4953–4957.

[9] C. Li and R. C. Hendriks, "Noise PSD insensitive RTF estimation in a reverberant and noisy environment," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech Signal Process.*, Jun. 2023, pp. 1–5.

[10] 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.

[11] A. Napolitano, "A — Nonstationary signal analysis," in *Cyclostationary Processes and Time Series*, A. Napolitano, Ed. Academic Press, Cambridge, MA, USA Jan. 2020, pp. 471–478. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780081027080000285

[12] G. Giannakis, "Cyclostationary signal analysis," in *Digital Signal Processing Fundamentals*, vol. 20094251. Boca Raton, FL, USA: CRC Press, Nov. 2009, pp. 1–32.

[13] Y. Liu, Z. Tan, H. Hu, L. J. Cimini, and G. Y. Li, "Channel estimation for OFDM," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1891–1908, Fourthquarter 2014.

[14] J. Wolfe, M. Garnier, N. H. Bernardoni, and J. Smith, "The mechanics and acoustics of the singing voice: Registers, resonances and the source–filter interaction," in *The Routledge Companion to Interdisciplinary Studies in Singing, Volume I: Development*. Evanston, IL, USA: Routledge, 2020.

[15] B. Dong, "Characterizing resonant component in speech: A different view of tracking fundamental frequency," *Mech. Syst. Signal Process.*, vol. 88, pp. 318–333, May 2017.

[16] H. Gockel, B. C. J. Moore, and R. D. Patterson, "Asymmetry of masking between complex tones and noise: The role of temporal structure and peripheral compression," *J. Acoust. Soc. Amer.*, vol. 111, no. 6, pp. 2759–2770, Jun. 2002.

[17] B. Moore, *An Introduction to the Psychology of Hearing*. Castle Hill, Australia: Emerald, 2012. [Online]. Available: https://books.google.nl/books?id=LM9U8e28pLMC

[18] C. Trahiotis and R. M. Stern, "Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1285–1293, Oct. 1989, doi: 10.1121/1.398743.

[19] I. A. McCowan, "Robust speech recognition using microphone arrays," Ph.D. dissertation, Queensland University of Technology, Brisbane, QLD, Australia, 2001.

[20] J. Dmochowski, J. Benesty, and S. Affes, "On spatial aliasing in microphone arrays," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1383–1395, Apr. 2009. [Online]. Available: https://ieeexplore.ieee.org/document/4696048/

[21] 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.

[22] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 544–548.

[23] 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. 26th Eur. Signal Process. Conf.*, Sep. 2018, pp. 2499–2503.

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

[25] J. Benesty, J. Chen, and E. A. P. Habets, "The bifrequency spectrum in speech enhancement," in *Speech Enhancement in the STFT Domain* (SpringerBriefs in Electrical and Computer Engineering Series), J. Benesty, J. Chen, and E. A. Habets, Eds. Berlin, Germany: Springer, 2012, pp. 93–101.

[26] H. Huang, L. Zhao, J. Chen, and J. Benesty, "A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction," *Digit. Signal Process.*, vol. 33, pp. 169–179, Oct. 2014.

[27] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[28] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 563–575, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9944916

[29] J. Antoni, G. Xin, and N. Hamzaoui, "Fast computation of the spectral correlation," *Mech. Syst. Signal Process.*, vol. 92, pp. 248–277, Aug. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327017300134

[30] G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins Studies in the Mathematical Sciences Series), 4th ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 2013.

[31] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 11, Dec. 2003, Art. no. 495250. [Online]. Available: https://asp-eurasipjournals.springeropen.com/articles/10.1155/S111086570330602X

[32] R. R. Nadakuditi and J. W. Silverstein, "Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 3, pp. 468–480, Jun. 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5447639/

[33] E. Anderson et al., *LAPACK Users' Guide*, 3rd ed. Philadelphia, PA, USA: SIAM, 1999.

[34] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1783–1795, Oct. 1990.

[35] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[36] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[37] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc. F. Commun., Radar Signal Process.*, vol. 130, no. 1, pp. 11–16, Feb. 1983. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/ip-f-1.1983.0003

[38] A. van den Bos, "A Cramer-Rao lower bound for complex parameters," *IEEE Trans. Signal Process.*, vol. 42, no. 10, p. 2859, Oct. 1994.

[39] D. G. Altman and J. M. Bland, "Standard deviations and standard errors," *Brit. Med. J.*, vol. 331, no. 7521, Oct. 2005, Art. no. 903.

[40] H. Kasasbeh, R. Viswanathan, and L. Cao, "Noise correlation effect on detection: Signals in equicorrelated or autoregressive(1) Gaussian," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1078–1082, Jul. 2017.

[41] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd Annu. ACM Conf. Multimedia.*, 2015, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[42] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 313–317.

[43] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[44] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.
[45] A. F. Molisch, *Wireless Communications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.
[46] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: Half a century of research," *Signal Process.*, vol. 86, no. 4, pp. 639–697, 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165168405002409

Giovanni Bologni received the B.Sc. degree in electronics engineering from the Polytechnic University of Turin, Turin, Italy, in 2017, and the M.Sc. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2020. From 2018 to 2020, he was with Sony R&D Stuttgart, Germany, firstly as an Intern Student, and later as a Development Engineer. In 2021, he was a Research Assistant with the Sapienza University of Rome, Roma, Italy. Since 2022, he has been working as a Ph.D. with the Signal Processing Systems group, Delft University of Technology. His research focuses on statistical signal processing and machine learning for speech and audio applications, with an emphasis on multichannel speech enhancement.

Richard C. Hendriks (Senior Member, IEEE) was born in Schiedam, The Netherlands. He received the B.Sc., M.Sc. (cum laude), and Ph.D. (cum laude) degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively. He is currently an Associate Professor with the Signal Processing Systems group, Delft University of Technology, and the Director of Studies of B.Sc. Electrical Engineering program. He was an Associate Editor for both the IEEE/ACM Transaction on Audio, Speech, and Language Processing from 2015–2019 and the EURASIP Journal on Advances in Signal Processing from 2015–2020. He is currently a Senior Associate Editor with the IEEE/ACM Trans. on Audio, Speech, and Language Processing. He was with IEEE as an Elected Member of IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing during two terms: a first term during the years 2019–2021 and a second term starting in 2024. In addition, he was one of the main organisers and Financial Chair of Eusipco 2020 in Amsterdam. His research interests include biomedical signal processing, audio and speech processing including speech enhancement, speech intelligibility improvement and intelligibility modelling. In March 2010, he received a VENI grant for his proposal "Intelligibility Enhancement for Speech Communication Systems". He was the recepient of several best paper awards, among which the IEEE Signal Processing Society best paper award in 2016.

Richard Heusdens(Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively. Since 2002, he has been an Associate Professor in the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. In 1992, he joined the digital signal processing group with Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group of Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio/speech signal processing activities within the ICT group. He held visiting positions at KTH (Royal Institute of Technology, Sweden) in 2002 and 2008 and was a Guest Professor with Aalborg University from 2014-2016. Since 2019, he is a Full Professor with the Netherlands Defence Academy. His research interests include audio and acoustic signal processing, sensor signal processing, distributed optimization and security/privacy.