IEEE Open Journal of Signal Processing

Received 17 October 2025; revised 12 January 2026; accepted 16 January 2026. Date of publication 26 January 2026; date of current version 13 February 2026. The review of this article was arranged by Associate Editor Dan Oneata.

Digital Object Identifier 10.1109/OJSP.2026.3657698

From Informed Independent Vector Extraction to Hybrid Architectures for Target Source Extraction

ZBYNĚK KOLDOVSKÝ [1] (Senior Member, IEEE), JIŘÍ MÁLEK [1] (Member, IEEE), MARTIN VRÁTNÝ [1], TEREZA VRBOVÁ [1], JAROSLAV ČMEJLA [1], AND STEPHEN O'REGAN [2]

[1] Acoustic Signal Analysis, Processing Group, Technical University of Liberec, 461 17 Liberec, Czech Republic
[2] Naval Surface Warfare Center Carderock Division, West Bethesda, MD 20817 USA

CORRESPONDING AUTHOR: ZBYNĚK KOLDOVSKÝ (email: zbynek.koldovsky@tul.cz).

This work was supported in part by Czech Science Foundation (GAČR) under Project25-18485S, in part by the US Office of Naval Research Global Project under Grant N62909-23-1-2084, and in part by the Ministry of Education, Youth and Sports of the Czech Republic.

ABSTRACT This article revises informed independent vector extraction (iIVE) as a framework for connecting model-based blind source extraction (BSE) with deep learning. We introduce the contrast function for iIVE, which is derived by extending IVE with beamforming-based constraints, enabling an interpretable use of reference signals. We also show that structured mixing models implementing physical knowledge can be integrated, which is demonstrated by two far-field models. With the contrast functions, rapidly converging second-order algorithms are developed, whose performance is first verified through simulations. In the experimental part, we refine iIVE by training models containing unrolled iterations of the developed algorithm. The resulting structures achieve performance comparable to state-of-the-art networks while requiring two orders of magnitude fewer trainable parameters and exhibiting strong generalization to unseen conditions.

INDEX TERMS Blind source separation, blind source extraction, independent component analysis, independent vector analysis, speaker extraction.

I. INTRODUCTION

A. MOTIVATION

Extracting a signal of interest (SOI) from multichannel noisy observations has been a fundamental task in signal processing and machine learning. The key to solving this problem is information directly about or related to the SOI. Various approaches attempt to capture this information through mathematical assumptions and training. It is very important to have methods whose requirements for applicability in different situations are as unrestrictive as possible.

Therefore, the defining philosophy of Blind Signal Extraction (BSE) has been that methods should only use information contained in the observed data [1]. In contrast, modern deep learning (DL) methods assume a sufficient number of examples to train nonlinear systems that are capable of solving the task in general situations. While BSE is limited by the available data and therefore in accuracy and uncertainty, DL methods are difficult to interpret, have unclear generalizability to unseen scenarios, and often face high computational demands. The current trend is therefore moving towards combining the advantages of these two and other approaches.

This article focuses on informed independent vector extraction (iIVE) as a suitable means of combining BSE, physical models, and DL. We review the theoretical basis for deriving the contrast function for iIVE, starting from IVE and introducing beamforming-based parameter constraints that allow an interpretable incorporation of reference signals. We show that the contrast function can also be combined with structured mixing models implementing physical knowledge and derive rapidly converging second-order algorithms. Attention is paid to the experimental part, where, in addition to verifying the algorithms by simulations, we optimize modules for obtaining side information through an unrolled algorithm. From the foundations of iIVE, we thus move towards hybrid architectures that can combine three sources of information:

physical knowledge (structured mixing models), information-theoretical principles (independence-based IVE), and side information obtained by a trained part.

B. BACKGROUND

The most successful methods for Blind Source Separation (BSS)[1] include Independent Component Analysis (ICA) and Nonnegative Matrix Factorization (NMF) [3], [4]. Later, extensions appeared for the joint separation of multiple mixtures: Independent Vector Analysis (IVA) and Multichannel NMF [5], [6]. This work is focused on methods following ICA.

BSS generally aims to retrieve all original signals contained in the observed mixture, while BSE is a special case of BSS in this respect, where we want to extract only the SOI [7], a group of signals [8], or their subspace [9]. Independent Component Extraction (ICE) and IVE are BSE variants of ICA and IVA, respectively, where the principle for identifying the SOI is its statistical independence from other signals in the mixture [10].

The uncertainty of the order of the original signals and their scale is a natural property of BSS and, in general, of all problems where information about target outputs is missing [11]. The uncertainties give rise to problems that are crucial to solve in practical deployments, such as the permutation problem [12], the discontinuity problem [13], and scaling ambiguity [14]. For example, to apply BSE in speaker extraction, we need to identify which signals belong to the target speaker, noise, and the other speakers.

For purely blind methods (BSS, BSE), it is necessary to assume that additional information is available and, depending on its type, define a strategy for using it. This is what informed methods [15], also called semi-blind [16] or guided [17], [18], attempt to do. Given the various forms of additive information and applications, these methods constitute a diverse set of approaches that may be similar or even equivalent, but may also be useful only for specific tasks. To name some, there are methods trying to control convergence through constraints imposed on parameters [19], [20], [21], [22] or directly on the extracted signal [23], [24]. Bayesian methods exploiting a priori probabilistic models can lead to similar solutions [25]. Incorporation of a reference featuring dependency with the SOI has been considered, e.g., in [17], [26], [27], [28].

With the advent of DL, attention has focused on training deep networks to solve separation and extraction tasks. The most widely used multichannel approaches aim to estimate time-frequency masks indicating the activity of (target) signals, which are then used to calculate covariance matrices and beamformers [29], [30]. Due to poor generalizability to unseen scenarios and weak interpretability, ways to combine DNN with knowledge from beamforming and blind signal separation were soon sought. In the field of target speaker extraction (TSE), the main problem addressed was how to identify the target person. For this purpose, it is possible to use embeddings derived from reference utterances of a specific person (enrollment) [31]. Spatial information [32] or even multimodal information [33] can also be used. A recent study compares the effectiveness of spatial information with embeddings [34]. Other recent approaches include generative models, such as diffusion architectures [35].

The idea that informed algorithms can appropriately combine BSS and DL is becoming increasingly common [27], [36]. The main focus is on optimizing source models in methods derived from IVA/IVE [37], [38], [39]. For example, DNN-based source power spectra estimation is considered in [40], [41]. A trainable surrogate function in auxiliary function-based IVA was proposed in [42], and its combination with the original Laplacean model from [43] was proposed in [44]. The trained source model was combined with geometrical constraints in [45]. An end-to-end automatic speech recognition system endowed by a speech separation method with a neural source model was considered in [46]. There are countless ways in which BSS and DL methods can be combined, so further work on this topic can be expected in the near future.

C. CONTRIBUTION

This article focuses on iIVE and its applicability in trainable hybrid models. The contribution has four main parts. First, we revise iIVE by introducing a suitable contrast function that involves parameter constraints employing reference signals. Second, we show that the contrast function can also be used with structured mixing models that reflect physical knowledge. Third, we show that the contrast functions can be used to develop second-order algorithms similar to the well-known FastICA [47], specifically, for the basic unstructured and two structured mixing models. Fourth, we verify the algorithms on simulated data and evaluate their involvement in showcase hybrid models for TSE.

The source extraction problem is formulated in Section II, where we introduce the mixing and de-mixing model parameterizations, including two structured models for linear sensor arrays. In Section III, we take the likelihood function for IVE and modify it by constraining the mixing and de-mixing parameters. The constraint is a function of the reference signals and is intuitively related to the minimum variance distortionless (MVDR) beamformer, which is well-known in array processing theory [48]. Based on this, we also define the contrast functions for two structured mixing models. The models assume linear sensor arrays and non-reflective environments, which are implemented in nonlinear parameterizations of the mixing vector with a smaller number of parameters than in the unstructured model.

In Section IV, the gradient and Hessian matrix of the contrast function of the unstructured model are derived in closed

[1] In this article, we use the term "blind" to refer to methods that use only observed data as input, and no training data. This definition is closer to the original idea from the 1990s [1] and differs from the recent trend, which includes a group of trained models [2]. The reason for this is the distinction between purely mathematical and data-driven models, combinations of which give rise to hybrid models, which are the focus of this article.

form. We show that they provide useful tools for developing second-order informed FastICA-like algorithms. For the unstructured model, we obtain a method that is very similar to the proven algorithms from [49], [50]. The framework presented here provides new insight into the necessity of using constraints that ensure subspace orthogonality. Moreover, we confirm the validity of the framework by deriving new informed algorithms also for the structured models. One of these algorithms corresponds to a novel informed variant of the single-parameter algorithm from [51]. We verify the algorithms by simulations in Section V, by which the relevance of the proposed contrast functions for iIVE is confirmed.

In the experimental Section VI, we verify the applicability of methods in trainable hybrid models for TSE. Reverberant mixtures of speakers are considered such that the target speaker position is within a defined sector of the room. First, a noise-only activity detector (NAD) is realized by a small-scale DNN that is pre-trained on a training subset. Then, the proposed algorithms are tested when the reference signal is obtained as the NAD output. Secondly, the pre-trained NAD is optimized in a hybrid architecture, where NAD is connected as a reference input to five iterations of the iIVE algorithm. In further experiments, we change the parameters of mixtures, such as reverberation time and room dimensions, and evaluate the ability of methods to generalize to unseen scenarios and compare their performance with state-of-the-art TSE models.

Codes of methods, simulations, and experiments are publicly available.[2]

C. NOMENCLATURE

Plain, bold, and bold capital letters denote scalars, vectors, and matrices, respectively. Upper index $\cdot^T$, $\cdot^H$, or $\cdot^*$ denotes, respectively, transposition, conjugate transpose, or complex conjugate. The Matlab convention for matrix/vector concatenation will be used, e.g., $[1;\ \mathbf{g}] = [1,\ \mathbf{g}^T]^T$. We will consider complex-valued signals and parameters.

II. PROBLEM FORMULATION

A. OBSERVED DATA

Let us consider K signal mixtures observed by multiple sensors. The kth mixture is assumed to obey the linear mixing model

$$\mathbf{x}_k(n) = \mathbf{a}_k s_k(n) + \mathbf{y}_k(n), \tag{1}$$

where n is the sample index $n = 1, \ldots, N$; $\mathbf{x}_k(n)$ is the $d \times 1$ vector of observed signals in the kth mixture, and $s_k(n)$ and $\mathbf{y}_k(n)$ denote, respectively, the SOI and the other signals; $\mathbf{a}_k$ is the $d \times 1$ *mixing vector* whose elements correspond to weights with which $s_k(n)$ is observed in the mixture. The goal is to retrieve $s_k(n)$ from each mixture.

As an example, k may correspond to a frequency in Short-term Fourier Transform (STFT) of audio time-domain signals recorded by d microphones, where $\mathbf{a}_k$ corresponds to acoustic or relative transfer functions of the signal paths between the target speaker and the microphones [52]. Similarly, k can be the index of a dataset, snapshot, or subject in a multi-subject experiment with biomedical data [53]. If $K > 1$, we are talking about joint extraction, where multiple mixtures are processed simultaneously.

There are indeterminacies inherent to the problem: $\mathbf{a}_k$ and $s_k(n)$ can have arbitrary scale as $(\delta\mathbf{a}_k)(\frac{1}{\delta}s_k(n)) = \mathbf{a}_k s_k(n)$ for any $\delta \neq 0$, which is referred to as the scaling ambiguity. Also, the role of $s_k(n)$ can be played by any other signal in the mixture that satisfies the properties by which we determine the SOI; this is referred to as the SOI uncertainty; see also Section II.B in [10].

In addition to the mixtures, we assume that scalar reference signals $r_k(n)$ are available, which might carry side information about the SOI; in case they do not depend on k, we denote the sole reference signal by $r(n)$. The primary purpose of the reference signals is to deal with the SOI uncertainty. They allow for the inclusion of various types of supplementary information (e.g., voice/noise activity detection, speaker ID, video, spatial information, etc.).

B. MIXING MODEL PARAMETERIZATION

Let $\mathbf{w}_k$ be the *separating vector*[3] that should extract the SOI from the kth mixture as $s_k(n) = \mathbf{w}_k^H \mathbf{x}_k(n)$. The existence of the ideal separating vector such that extracts $s_k(n)$ without any distortion and residual interference is not guaranteed by (1) in general. We now introduce assumptions that guarantee the existence of the ideal $\mathbf{w}_k$, under which the model is referred to as determined.

Assume that $\mathbf{y}_k(n)$ only span a $(d-1)$-dimensional subspace of so-called background signals, which will be denoted by a $(d-1) \times 1$ vector $\mathbf{z}_k(n)$. It means that a $(d-1) \times d$ blocking matrix $\mathbf{B}(\mathbf{a}_k)$ (a function of $\mathbf{a}_k$) exists such that the background signals can be obtained through $\mathbf{z}_k(n) = \mathbf{B}(\mathbf{a}_k)\mathbf{x}_k(n)$. Let $\mathbf{a}_k$ be divided as $\mathbf{a}_k = [\gamma_k; \mathbf{g}_k]$ and let us define $\mathbf{B}(\mathbf{a}_k) = [\mathbf{g}_k \quad -\gamma_k \mathbf{I}_{d-1}]$. It holds that $\mathbf{B}(\mathbf{a}_k)\mathbf{a}_k = \mathbf{0}$. Thus, $\mathbf{B}(\mathbf{a}_k)$ is a blocking matrix and the background signals can be defined as $\mathbf{z}_k(n) = \mathbf{B}(\mathbf{a}_k)\mathbf{x}_k(n) = \mathbf{B}(\mathbf{a}_k)\mathbf{y}_k(n)$.

Now, the de-mixing model (the inverse of the mixing model) can be described by a non-singular square de-mixing matrix $\mathbf{W}_k$ where it holds that

$$\mathbf{W}_k \mathbf{x}_k(n) = \begin{bmatrix} \mathbf{w}_k^H \\ \mathbf{B}(\mathbf{a}_k) \end{bmatrix} \mathbf{x}_k(n) = \begin{bmatrix} s_k(n) \\ \mathbf{z}_k(n) \end{bmatrix}. \tag{2}$$

In hindsight, the mixing system (1) can be therefore described by a mixing matrix $\mathbf{A}_k = \mathbf{W}_k^{-1}$, where

$$\mathbf{x}_k(n) = \mathbf{A}_k \begin{bmatrix} s_k(n) \\ \mathbf{z}_k(n) \end{bmatrix}. \tag{3}$$

[2] https://github.com/ASAP-Group/hybridIVE.git

[3] In array processing theory, the separating vector $\mathbf{w}_k$ would be called a beamformer.

By dividing the separating vector as $\mathbf{w}_k = [\beta_k; \mathbf{h}_k]$, we can express the analytic form of $\mathbf{A}_k = \mathbf{W}_k^{-1}$

$$\mathbf{A}_k = \begin{bmatrix} \beta_k^* & \mathbf{h}_k^H \\ \mathbf{g}_k & -\gamma_k \mathbf{I}_{d-1} \end{bmatrix}^{-1} = \begin{bmatrix} \gamma_k & \mathbf{h}_k^H \\ \mathbf{g}_k & \frac{1}{\gamma_k}(\mathbf{g}_k \mathbf{h}_k^H - \mathbf{I}_{d-1}) \end{bmatrix}, \quad (4)$$

which is valid (the reader can verify that $\mathbf{W}_k \mathbf{A}_k = \mathbf{I}_d$) under the *distortionless* condition that

$$\mathbf{w}_k^H \mathbf{a}_k = \beta_k^* \gamma_k + \mathbf{h}_k^H \mathbf{g}_k = 1. \quad (5)$$

To summarize, we assume that, besides (1), the observed data also obey (3) where $\mathbf{A}_k$ is square and is parametrized by the parametric vectors $\mathbf{a}_k$ and $\mathbf{w}_k$ according to (4), and $\mathbf{a}_k$ and $\mathbf{w}_k$ satisfy (5). The source extraction problem can now be formulated to find $\mathbf{w}_k$. However, in the blind and semi-blind settings considered in this paper, we will see that this also implies the finding of $\mathbf{a}_k$.

The mixing models with the square non-singular mixing matrix has been referred to as *determined* [1], [37], [54]. It enables us to apply the classical statistical estimation as the true values of $\mathbf{a}_k$ and $\mathbf{w}_k$ are assumed to exist. This assumption does not necessarily mean a limitation in practice. In the experimental section, we also consider situations where the mixtures do not obey the model perfectly.

In this article, we will specifically address the following special cases.

1) UNSTRUCTURED MIXING MODEL

As a basis for further methods, we will consider the conventional mixing model where no additional structure is assumed for the parametric vectors $\mathbf{a}_k$ and $\mathbf{w}_k$.

2) PHASE-SHIFT MIXING MODEL

Of particular interest will be the case where the mixing vector is structured according to

$$\mathbf{a}_k(\lambda_k) = \begin{bmatrix} 1 & e^{\mathrm{i}\lambda_k v_{2,k}} & \dots & e^{\mathrm{i}\lambda_k v_{d,k}} \end{bmatrix}^T, \quad (6)$$

where λ_k are real-valued, and $v_{j,k}$, $j = 2, \dots, d$, are known weights collected in the vectors $\mathbf{v}_k = [0, v_{2,k}, \dots, v_{d,k}]^T$, $k = 1, \dots, K$; i denotes the imaginary unit.

When the vectors $\mathbf{v}_k$ are real-valued, the model describes situations where the contribution of the signal $s_k(n)$ is phase-shifted on input channels while the magnitude is the same. A typical case is the impact of a plane wave on a linear array of sensors at an angle whose cosine is given by λ_k. The phase shifts are then given by

$$v_{m,k} = \frac{2\pi f_s}{c_k} \frac{k-1}{2(K-1)}(m-1)d_m, \quad (7)$$

where f_s is the sampling frequency, $2(K-1)$ is the window length of the STFT, c_k is the speed of signal propagation, and d_m is the distance of the mth sensor from the $(m-1)$th sensor [55].

3) FAR-FIELD MIXING MODEL

In the previous model, the angle at which the planar wave is traveling may be frequency-dependent. The special case considered here is when

$$\lambda = \lambda_1 = \dots = \lambda_K, \quad (8)$$

that is when the angle is the same for all frequencies. In acoustics, this is the case when the sound source (e.g. a speaker) is a point source and is sufficiently distant from the microphone array, and the environment is free of any reflections (free-field).

III. CONTRAST FUNCTIONS

This section introduces contrast functions that enable the identification of $\mathbf{a}_k$ and $\mathbf{w}_k$. Contrast functions are derived based on the maximum likelihood principle, and parameter constraints play a special role here. In particular, the contrasts for the semi-blind estimation are derived through constraints employing the reference signals.

A. STATISTICAL MODEL OF DATA

Following the basic statistical model of the IVE, each signal is modeled as a sequence of identically and independently distributed zero-mean random variables [5], [10], [56]. We will denote the random variables using the same symbol as the corresponding samples but without the argument n. For instance, s_k will represent $s_k(n)$, $n = 1, \dots, N$.

The key assumption in IVE is that the SOI s_k and the background signals $\mathbf{z}_k$ are mutually independent. The reference signals r_k are assumed to show dependencies with s_k but are independent of the background signals. Another key feature in IVE is that possible dependencies among $s_1, \dots, s_K$, i.e. of the SOIs in the K mixtures, are taken into account. Let $\mathbf{s} = [s_1, \dots, s_K]^T$, which is referred to as the SOI vector component. The dependencies are implemented through a joint non-Gaussian pdf $p(\mathbf{s}) = p(s_1, \dots, s_K)$, which need not be just the product of marginal pdfs of $s_1, \dots, s_K$ [5].

The background signals $\mathbf{z}_k$ are assumed to be zero-mean circular Gaussian having an unknown covariance matrix denoted by $\dot{\mathbf{C}}_{\mathbf{z}_k}$; $\mathbf{z}_{k_1}$ and $\mathbf{z}_{k_2}$ are assumed uncorrelated for $k_1 \neq k_2$. These assumptions mean that we neglect higher-order statistics and inter-mixture dependencies of the background signals, which often leads to a statistical suboptimality [57]. Nevertheless, this is worthwhile in terms of simplification; see, e.g., discussions in [13].

A. CONVENTIONS

From now on, we will distinguish the true values of signals and parameters by the dot accent. For example, $\dot{\mathbf{a}}_k$ and $\dot{\mathbf{w}}_k$ will, respectively, denote the true mixing and separating vectors. Similarly, $\dot{s}_k$ and $\dot{\mathbf{z}}_k$ denote the true SOI and the background signals, respectively. For simplicity, we will denote sample-based averages by the expectation operator $\mathrm{E}[\cdot]$, the value of which corresponds to the true expectation when $N \to +\infty$.

B. CONTRASTS FOR BLIND ESTIMATION

The contrast function for the blind estimation of $\mathbf{a}_k$ and $\mathbf{w}_k$, $k = 1, \ldots, K$, considering the unstructured mixing model is given by[4]

$$\mathcal{C}(\mathbf{w}_1, \mathbf{a}_1, \ldots, \mathbf{w}_K, \mathbf{a}_K) = \mathrm{E}\left[\log f(\bar{\mathbf{s}})\right] - \sum_{k=1}^{K} \log \sigma_k^2 - \sum_{k=1}^{K} \mathrm{E}\left[\mathbf{z}_k^H \mathbf{C}_{\mathbf{z}_k}^{-1} \mathbf{z}_k\right] + (d-2) \sum_{k=1}^{K} \log |\gamma_k|^2, \tag{9}$$

where σ_k^2 denotes the estimated variance of s_k, $\bar{\mathbf{s}} = [\frac{s_1}{\sigma_1}, \ldots, \frac{s_K}{\sigma_K}]^T$, and $f(\cdot)$ is a suitable model non-Gaussian pdf of normalized random variables; $f(\cdot)$ is needed to replace the unknown true pdf of $\mathbf{s}$.

In the case of the phase-shift model (6), the contrast function is readily given by

$$\mathcal{C}_{\mathbf{w},\boldsymbol{\lambda}}(\mathbf{w}_1, \lambda_1, \ldots, \mathbf{w}_K, \lambda_K) = \mathcal{C}(\mathbf{w}_1, \mathbf{a}_1(\lambda_1), \ldots, \mathbf{w}_K, \mathbf{a}_K(\lambda_K)) \tag{10}$$

and, for the far-field model (8), it is given by

$$\mathcal{C}_{\mathbf{w},\lambda}(\mathbf{w}_1, \ldots, \mathbf{w}_K, \lambda) = \mathcal{C}(\mathbf{w}_1, \mathbf{a}_1(\lambda), \ldots, \mathbf{w}_K, \mathbf{a}_K(\lambda)). \tag{11}$$

C. CONSTRAINTS

The parametric vectors $\mathbf{a}_k$ and $\mathbf{w}_k$ are not completely free variables due to the distortionless constraint (5). However, this link is often shown to be insufficient due to spurious local extremes of the contrast functions. It thus turns out to be advantageous to introduce stronger constraints, also for the sake of reducing the number of free variables.

1) MPDR

The so-called orthogonal constraint (OC) requires that the subspace generated by the current estimate of s_k is orthogonal to that of $\mathbf{z}_k$. The condition means that $\mathrm{E}[\mathbf{z}_k s_k^*] = \mathbf{B}(\mathbf{a}_k)\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_k = \mathbf{0}$, where $\mathbf{C}_{\mathbf{x}_k}$ is the sample covariance matrix $\mathbf{C}_{\mathbf{x}_k} = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$. The OC is consistent with the model since $\mathrm{E}[\dot{\mathbf{z}}_k s_k^*]$ is zero when $N \to +\infty$ because of the independence of s_k and $\dot{\mathbf{z}}_k$.

The OC is given by [10]

$$\mathbf{w}_{\mathrm{OC},k}(\mathbf{a}_k) = \frac{\mathbf{C}_{\mathbf{x}_k}^{-1}\mathbf{a}_k}{\mathbf{a}_k^H \mathbf{C}_{\mathbf{x}_k}^{-1}\mathbf{a}_k} = \sigma_k^2 \mathbf{C}_{\mathbf{x}_k}^{-1}\mathbf{a}_k, \tag{12}$$

when $\mathbf{w}_k$ is expressed as the dependent variable on $\mathbf{a}_k$. Conversely,

$$\mathbf{a}_{\mathrm{OC},k}(\mathbf{w}_k) = \frac{\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_k}{\mathbf{w}_k^H \mathbf{C}_{\mathbf{x}_k}\mathbf{w}_k} = \sigma_k^{-2}\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_k \tag{13}$$

when $\mathbf{a}_k$ is treated as the dependent variable.

[4]For a detailed justification of the contrast function, we refer the reader to Section III.B in [13] where the considered model coincides with the one here when $T = 1$ (static mixing model).

Note that (12) has the same analytic form as the minimum power distortionless beamformer (MPDR). To verify the orthogonality, note that $\mathbf{B}(\delta\mathbf{a})\mathbf{a} = \mathbf{0}$ for any vector $\mathbf{a}$ and scalar δ. Hence, $\mathrm{E}[\mathbf{z}_k s_k^*] = \mathbf{B}(\mathbf{a}_{\mathrm{OC},k}(\mathbf{w}_k))\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_k = \mathbf{B}(\mathbf{a}_k)\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\mathrm{OC},k} = \mathbf{0}$. Similarly, it is easy to verify that $\mathbf{w}_{\mathrm{OC},k}(\mathbf{a}_k)^H\mathbf{a}_k = \mathbf{w}_k^H\mathbf{a}_{\mathrm{OC},k}(\mathbf{w}_k) = 1$, which means that the couple $\mathbf{w}_k$ and $\mathbf{a}_{\mathrm{OC},k}(\mathbf{w}_k)$ as well as $\mathbf{w}_{\mathrm{OC},k}(\mathbf{a}_k)$ and $\mathbf{a}_k$ satisfy the distortionless condition. For $N \to +\infty$, $\mathbf{w}_{\mathrm{OC},k}(\dot{\mathbf{a}}_k) = \dot{\mathbf{w}}_k$ as well as $\mathbf{a}_{\mathrm{OC},k}(\dot{\mathbf{w}}_k) = \dot{\mathbf{a}}_k$.

The optimization of (9) under the OC has been widely studied and leads to the well-known ICE/IVE algorithms; see, e.g., [8], [10], [47].

2) MVDR

MPDR is known to be sensitive to the estimation errors in $\mathbf{C}_{\mathbf{x}_k}$ and $\mathbf{a}_k$. For example, $\mathbf{w}_{\mathrm{OC},k}(\mathbf{a}_k)$ can be significantly different from $\dot{\mathbf{w}}_k$ if $\mathbf{a}_k$ deviates too much from $\dot{\mathbf{a}}_k$ and/or when the estimation error in $\mathbf{C}_{\mathbf{x}_k}$ is large.

An alternative to MPDR is

$$\mathbf{w}_{\mathrm{MVDR},k}(\mathbf{a}_k) = \frac{\mathbf{C}_{\mathbf{y}_k}^{-1}\mathbf{a}_k}{\mathbf{a}_k^H \mathbf{C}_{\mathbf{y}_k}^{-1}\mathbf{a}_k}, \tag{14}$$

where $\mathbf{C}_{\mathbf{y}_k}$ is the sample covariance matrix of $\mathbf{y}_k$. For $N \to +\infty$, this is known as the minimum variance distortionless beamformer (MVDR). MVDR is less sensitive to the estimation error in $\mathbf{a}_k$ than MPDR. However, it requires the knowledge of $\mathbf{C}_{\mathbf{y}_k}$, which is hardly available in real situations, let alone in a blind scenario.

3) APPROXIMATE MVDR

In [49], we tried for the first time to replace the OC with an approximate MVDR where the unknown $\mathbf{C}_{\mathbf{y}_k}$ is replaced by a weighted covariance matrix

$$\mathbf{C}_{\alpha_k} = \mathrm{E}\left[\alpha_k \mathbf{x}_k \mathbf{x}_k^H\right]. \tag{15}$$

The corresponding constraint is

$$\mathbf{w}_{\alpha,k}(\mathbf{a}_k) = \frac{\mathbf{C}_{\alpha_k}^{-1}\mathbf{a}_k}{\mathbf{a}_k^H \mathbf{C}_{\alpha_k}^{-1}\mathbf{a}_k} = \sigma_{\alpha,k}^2 \mathbf{C}_{\alpha_k}^{-1}\mathbf{a}_k, \tag{16}$$

where α_k is a function of the reference signal r_k, referred to as the weighting function; $\sigma_{\alpha,k}^2 = (\mathbf{a}_k^H \mathbf{C}_{\alpha_k}^{-1}\mathbf{a}_k)^{-1} = \mathbf{w}_{\alpha,k}^H \mathbf{C}_{\alpha_k}\mathbf{w}_{\alpha,k}$. In general, the goal is to make $\widehat{\mathbf{C}}_{\alpha_k}$ as close to $\mathbf{C}_{\mathbf{y}_k}$ as possible. For example, α_k can be chosen as an indicator of the activity of the SOI. The consistency of (16) proves the following lemma.

Lemma 1: The separating vector estimate (16) is consistent in the sense that, for $N \to +\infty$, $\mathbf{w}_{\alpha,k}(\dot{\mathbf{a}}_k) = \dot{\mathbf{w}}_k$.

Proof: See Appendix A. ■

D. CONSTRAINT AND CONTRASTS FOR SEMI-BLIND ESTIMATION

In fact, (16) does not guarantee the orthogonality, i.e., $\mathbf{B}(\mathbf{a}_k)\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\alpha,k}(\mathbf{a}_k) \neq 0$ in general. We found that this turns out to be a problem in the algorithm development. Therefore, our

proposal in this paper is to combine the two constraints (12) and (16) by substituting each $\mathbf{a}_k$ and $\mathbf{w}_k$ according to

$$\mathbf{a}_k \leftarrow \mathbf{a}_{\alpha\mathrm{OC},k}(\mathbf{a}_k) = \mathbf{a}_{\mathrm{OC},k}\big(\mathbf{w}_{\alpha,k}(\mathbf{a}_k)\big) = \frac{\sigma^2_{\alpha,k}}{\varsigma^2_k}\mathbf{R}_k\mathbf{a}_k, \tag{17}$$

$$\mathbf{w}_k \leftarrow \mathbf{w}_{\alpha,k}(\mathbf{a}_k), \tag{18}$$

Since both substitutions are functions of $\mathbf{a}_k$, they represent a constraint alternative to (16). Here, we have introduced (for simplicity, we omit the arguments of $\mathbf{a}_{\alpha\mathrm{OC},k}$ and $\mathbf{w}_{\alpha,k}$ unless necessary)

$$\mathbf{R}_k = \mathbf{C}_{\mathbf{x}_k}\mathbf{C}^{-1}_{\alpha_k}, \tag{19}$$

$$\varsigma^2_k = \mathbf{w}^H_{\alpha,k}\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\alpha,k} = \sigma^4_{\alpha,k}\mathbf{a}^H_k\mathbf{C}^{-1}_{\alpha_k}\mathbf{C}_{\mathbf{x}_k}\mathbf{C}^{-1}_{\alpha_k}\mathbf{a}_k, \tag{20}$$

where ς^2_k corresponds to the sample variance of $s_k = \mathbf{w}^H_{\alpha,k}\mathbf{x}_k$.

The orthogonality and distortionless condition follow, respectively, from that $\mathrm{E}[\mathbf{z}_k s^*_k] = \mathbf{B}\left(\mathbf{a}_{\alpha\mathrm{OC},k}\right)\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\alpha,k} = \sigma^2_{\alpha,k}\mathbf{B}\left(\frac{\sigma^2_{\alpha,k}}{\varsigma^2_k}\mathbf{R}_k\mathbf{a}_k\right)\mathbf{R}_k\mathbf{a}_k = 0$ and

$$\mathbf{w}^H_{\alpha,k}\mathbf{a}_{\alpha\mathrm{OC},k} = \varsigma^{-2}_k\mathbf{a}^H_k\mathbf{C}^{-1}_{\alpha_k}\mathbf{R}_k\mathbf{a}_k = 1.$$

Lemma 2: For $N \to +\infty$, $\mathbf{a}_{\alpha\mathrm{OC},k}(\dot{\mathbf{a}}_k) = \dot{\mathbf{a}}_k$.

Proof: By Lemma 1, $\dot{\varsigma}^2_k = \mathbf{w}_{\alpha,k}(\dot{\mathbf{a}}_k)^H\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\alpha,k}(\dot{\mathbf{a}}_k) = \dot{\mathbf{w}}^H_k\mathbf{C}_{\mathbf{x}_k}\dot{\mathbf{w}}_k = \dot{\sigma}^2_k$. Hence,

$$\mathbf{a}_{\alpha\mathrm{OC},k}(\dot{\mathbf{a}}_k) = \frac{\dot{\sigma}^2_{\alpha,k}}{\dot{\sigma}^2_k}\mathbf{C}_{\mathbf{x}_k}\mathbf{C}^{-1}_{\alpha,k}\dot{\mathbf{a}}_k = \dot{\sigma}^{-2}_k\mathbf{C}_{\mathbf{x}_k}\dot{\mathbf{w}}_k = \dot{\mathbf{a}}_k. \tag{21}$$

■

New contrast functions for semi-blind estimation are obtained through employing (17)-(18) in (9)-(11). For the basic unstructured mixing model, the contrast function, which is a function of $\mathbf{a}_1, \ldots, \mathbf{a}_K$, will be

$$\mathcal{C}_{\mathbf{a}}(\mathbf{a}_1, \ldots, \mathbf{a}_K) = \mathcal{C}\big(\mathbf{w}_{\alpha,1}, \mathbf{a}_{\alpha\mathrm{OC},1}, \ldots, \mathbf{w}_{\alpha,K}, \mathbf{a}_{\alpha\mathrm{OC},K}\big). \tag{22}$$

Hence, for the phase-shift model, the contrast is

$$\mathcal{C}_{\boldsymbol{\lambda}}(\lambda_1, \ldots, \lambda_K) = \mathcal{C}_{\mathbf{a}}(\mathbf{a}_1(\lambda_1), \ldots, \mathbf{a}_K(\lambda_K)), \tag{23}$$

and for the far-field model,

$$\mathcal{C}_{\lambda}(\lambda) = \mathcal{C}_{\mathbf{a}}(\mathbf{a}_1(\lambda), \ldots, \mathbf{a}_K(\lambda)), \tag{24}$$

which is a function of a single real-valued parameter. It is worth noting here that (23) and (24) offer alternatives to regularized contrasts used, e.g., in [21], [22]. While the former considers the specific structure of mixing parameters, the latter only enforces proximity to these structures.

IV. ALGORITHMS

We now derive algorithms based on optimizing the above contrast functions. We start with a detailed computation of the gradient and Hessian matrices of the contrast function (22) for the unstructured model. The results will be important for the further development of algorithms.

A. DERIVATIVES FOR UNSTRUCTURED MODEL

The gradient and the Hessian matrices of (22) and the assumptions under which they are computed are formulated in the following two statements.

Statement 1: Let the unknown score functions $-\frac{\partial \log f(\mathbf{s})}{\partial s_k}$, $k = 1, \ldots, K$, be replaced by $\nu^{-1}_k\phi_k(\bar{\mathbf{s}})$, where $\phi_k(\mathbf{s})$ are suitable (scalar) nonlinear functions[5], $\bar{\mathbf{s}} = [\frac{s_1}{\varsigma_1}, \ldots, \frac{s_K}{\varsigma_K}]^T$, and

$$\nu_k = \mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\frac{s_k}{\varsigma_k}\right]. \tag{25}$$

Let the unknown covariance matrix $\dot{\mathbf{C}}_{\mathbf{z}_k}$ be replaced by the sample-based covariance of the current estimate of $\mathbf{z}_k$. Then, the gradient of (22) reads

$$\boldsymbol{\Delta}_k = \frac{\partial \mathcal{C}_{\mathbf{a}}}{\partial \mathbf{a}^*_k} = \sigma^2_{\alpha,k}\mathbf{C}^{-1}_{\alpha_k}\left(\mathbf{a}_{\alpha\mathrm{OC},k}(\mathbf{a}_k) - \nu^{-1}_k\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k}{\varsigma_k}\right]\right). \tag{26}$$

Proof: See Appendix B. ■

Statement 2: Let, for all $k = 1, \ldots, K$, $\mathbf{a}_k = \dot{\mathbf{a}}_k$, $N \to +\infty$, and ν_k be treated as constants. Define

$$\xi_k = \mathrm{E}\left[\frac{|s_k|^2}{\sigma^2_k}\frac{\partial\phi_k(\mathbf{s})}{\partial s^*_k}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\right], \tag{27}$$

$$\eta_k = \mathrm{E}\left[\frac{s^2_k}{\sigma^2_k}\frac{\partial\phi_k(\mathbf{s})}{\partial s_k}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\right], \tag{28}$$

$$\rho_k = \mathrm{E}\left[\frac{\partial\phi_k(\mathbf{s})}{\partial s^*_k}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\right], \tag{29}$$

$$\xi_{\ell,k} = \mathrm{E}\left[\frac{s^*_\ell}{\sigma_\ell}\frac{s_k}{\sigma_k}\frac{\partial\phi_k(\mathbf{s})}{\partial s^*_\ell}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\right], \tag{30}$$

$$\eta_{\ell,k} = \mathrm{E}\left[\frac{s_\ell}{\sigma_\ell}\frac{s_k}{\sigma_k}\frac{\partial\phi_k(\mathbf{s})}{\partial s_\ell}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\right]. \tag{31}$$

and assume that the nonlinearities $\phi_k(\cdot)$ are chosen such that

$$\xi_k - \eta_k - \nu_k = 0, \tag{32}$$

and

$$\xi_{\ell,k} = \eta_{\ell,k}. \tag{33}$$

Then,

$$\frac{\partial\boldsymbol{\Delta}^T_k}{\partial\mathbf{a}_k} = \frac{\dot{\nu}_k - \dot{\rho}_k}{\dot{\nu}_k}\left(\frac{\dot{\sigma}^4_{\alpha,k}}{\dot{\sigma}^2_k}\dot{\mathbf{C}}^{-*}_{\alpha_k}\dot{\mathbf{C}}^*_{\mathbf{x}_k}\dot{\mathbf{C}}^{-*}_{\alpha_k} - \dot{\mathbf{w}}^*_k\dot{\mathbf{w}}^T_k\right), \tag{34}$$

$$\frac{\partial\boldsymbol{\Delta}^H_k}{\partial\mathbf{a}_\ell} = \mathbf{0}, \qquad \ell = 1, \ldots, K, \tag{35}$$

$$\frac{\partial\boldsymbol{\Delta}^T_k}{\partial\mathbf{a}_\ell} = \mathbf{0}, \qquad \ell \neq k. \tag{36}$$

Proof: See Appendix C. ■

We will now derive second-order algorithms based on the Newton-Raphson (NR) scheme for complex-valued signals and parameters; see (28) in [58]. These methods involve the inversion of the Hessian matrix. An important simplification

[5]The scaling of the nonlinear functions is needed so that the gradient of (22) is zero when $N = +\infty$ and $\mathbf{a}_k = \dot{\mathbf{a}}_k$; see, e.g., Section III.A in [10].

here brings the assumption of Section III-A that the background signals are circular: Together with condition (32), it causes (35) to be equal to zero for $\ell = k$. For $K > 1$, (33) plays an important role, due to which the Hessian matrix is block-diagonal ((35) and (36) are zero for $\ell \neq k$). Thus, only the inverses of the diagonal blocks (34) need to be computed in order to evaluate the Newton-Raphson update rule.

Conditions (32) and (33) can be satisfied by a suitable choice of the nonlinear functions $\phi_k(\cdot)$. The following lemma offers one possible choice that we will consider further in this article.

Lemma 3: Rational nonlinear functions

$$\phi_k^{\text{rati}}(\mathbf{s}) = \frac{s_k^*}{1 + \sum_{j=1}^K |s_j|^2}, \quad k = 1, \ldots, K, \tag{37}$$

satisfy the conditions defined by (32) and (33).

Proof: The proof, which can be easily done using the definitions (25)-(31), is left to the reader. ■

B. INFORMED ALGORITHM FOR UNSTRUCTURED MODEL

Given an initial value of $\mathbf{a}_k$, $k = 1, \ldots, K$, denoted by $\mathbf{a}_k^{\text{ini}}$, we now employ the NR scheme to find a local maximum of (22). Which source the algorithm extracts depends on the initialization and the reference signal.

Nonlinearities such as (37) are assumed to meet the conditions of Statement 2. As the Hessian matrix becomes block diagonal, the updates of the mixing vectors $\mathbf{a}_1, \ldots, \mathbf{a}_K$ can be computed separately. The NR rule suggests the update $\mathbf{a}_k \leftarrow \mathbf{a}_k - \mathbf{H}_k^{-*}\boldsymbol{\Delta}_k$ where $\mathbf{H}_k$ is given by (34), that is

$$\mathbf{H}_k = \frac{\dot{\nu}_k - \dot{\rho}_k}{\dot{\nu}_k}\left(\frac{\dot{\sigma}_{\alpha,k}^4}{\dot{\sigma}_k^2}\dot{\mathbf{C}}_{\alpha_k}^{-*}\dot{\mathbf{C}}_{\mathbf{x}_k}^{*}\dot{\mathbf{C}}_{\alpha_k}^{-*} - \dot{\mathbf{w}}_k^{*}\dot{\mathbf{w}}_k^{T}\right). \tag{38}$$

However, in order to implement this procedure, two issues must be resolved:

1) The true values of the statistics that appear in (34) are not known. Therefore, we replace these values with the corresponding sample-based averages, with the true value of the SOI being replaced by its current estimate.
2) $\mathbf{H}_k$ is rank deficient, which the reader can verify by showing that $\mathbf{H}_k\dot{\mathbf{a}}_k^* = \mathbf{0}$. This property is caused by the scaling ambiguity (see Section II-A): the scale of $\dot{\mathbf{a}}_k$ can be arbitrary (and compensated by the scale of $\dot{\mathbf{w}}_k$ or $\dot{s}_k$), but the contrast function (22) does not constrain it in any way. An efficient solution is to remove the rank-one term from (34). We, therefore, introduce augmented[6] Hessian matrices

$$\widetilde{\mathbf{H}}_k = \frac{\nu_k - \rho_k}{\nu_k}\frac{\sigma_{\alpha,k}^4}{\varsigma_k^2}\mathbf{C}_{\alpha_k}^{-*}\mathbf{C}_{\mathbf{x}_k}^{*}\mathbf{C}_{\alpha_k}^{-*}. \tag{39}$$

Now, the NR scheme suggests updating each $\mathbf{a}_k$ as

$$\mathbf{a}_k \leftarrow \mathbf{a}_k - \widetilde{\mathbf{H}}_k^{-*}\boldsymbol{\Delta}_k$$

[6]A detailed rationale for this step is provided by Proposition 2 in [13].

Algorithm 1: The iFastIVE Algorithm.

Input: $\mathbf{x}_k(n)$, $\mathbf{a}_k = \mathbf{a}_k^{\text{ini}}$, $\alpha_k(n)$, tol, $n = 1, \ldots, N$
Output: $\mathbf{a}_k$, $\mathbf{w}_k$, $s_k(n)$

1 **repeat**
2 **for** $k = 1, \ldots, K$ **do**
3 $\mathbf{a}_{k,\text{old}} = \mathbf{a}_k$;
4 $\sigma_{\alpha,k}^2 = (\mathbf{a}_k^H\mathbf{C}_{\alpha_k}^{-1}\mathbf{a}_k)^{-1}$;
5 $\mathbf{w}_k = \sigma_{\alpha,k}^2\mathbf{C}_{\alpha_k}^{-1}\mathbf{a}_k$;
6 $\varsigma_k^2 = \mathbf{w}_k^H\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_k$;
7 $\mathbf{a}_k = \varsigma_k^{-2}\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_k$;
8 $s_k(n) = \mathbf{w}_k^H\mathbf{x}_k(n)$;
9 **end**
10 $\bar{\mathbf{s}}(n) = [\frac{s_1(n)}{\varsigma_1}, \ldots, \frac{s_K(n)}{\varsigma_K}]^T$;
11 **for** $k = 1, \ldots, K$ **do**
12 $\rho_k = \mathrm{E}\left[\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial s_k^*}\right]$;
13 $\mathbf{a}_k \leftarrow \mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k}{\sigma_k}\right] - \rho_k\mathbf{a}_k$;
14 **end**
15 crit $= \max_{k=1,\ldots,K}\left\{1 - \frac{|\mathbf{a}_k^H\mathbf{a}_{k,\text{old}}|}{\|\mathbf{a}_k\|\|\mathbf{a}_{k,\text{old}}\|}\right\}$;
16 **until** crit $<$ tol;

$$= \mathbf{a}_k - \frac{\nu_k}{\nu_k - \rho_k}\frac{\varsigma_k^2}{\sigma_{\alpha,k}^2}\mathbf{R}_k^{-1}\left(\mathbf{a}_{\alpha\text{OC},k} - \nu_k^{-1}\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k}{\sigma_k}\right]\right),$$

$$= \mathbf{a}_k - \frac{\nu_k}{\nu_k - \rho_k}\mathbf{a}_k + \frac{1}{\nu_k - \rho_k}\frac{\varsigma_k^2}{\sigma_{\alpha,k}^2}\mathbf{R}_k^{-1}\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k}{\sigma_k}\right] \tag{40}$$

We use the fact that $\mathbf{a}_k$ can be arbitrarily scaled and multiply the right-hand side (RHS) of (40) by $\nu_k - \rho_k$, which gives

$$\mathbf{a}_k \leftarrow \frac{\varsigma_k^2}{\sigma_{\alpha,k}^2}\mathbf{R}_k^{-1}\left(\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k}{\sigma_k}\right] - \rho_k\mathbf{a}_{\alpha\text{OC},k}\right). \tag{41}$$

In practice, we have found that it is important to multiply the RHS of (41) by $\frac{\sigma_{\alpha,k}^2}{\varsigma_k^2}\mathbf{R}_k$, which can be explained by the need to project the original value to the manifold specified by the constraint (17). The final update rule then simplifies to

$$\mathbf{a}_k \leftarrow \mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k}{\sigma_k}\right] - \rho_k\mathbf{a}_{\alpha\text{OC},k}, \tag{42}$$

which gives us a new starting value of $\mathbf{a}_k$ for the next iteration. The application of the constraints (17)-(18) and the update (42) are repeated until a stopping criterion reflecting the latest changes is $\mathbf{a}_k$, $k = 1, \ldots, K$ is satisfied (see line 15in Algorithm 1). The resulting algorithm will be referred to as iFastIVE; its pseudo-code is summarized in Algorithm 1.

1) PREVIOUS VARIANTS OF IFASTIVE

The iFastIVE algorithm itself is not particularly new. Similar methods have been recently derived in [49], [50]. However, the previous derivations, which did not utilize constraints (17)-(18), required heuristic steps that cannot be applied generally, for example, in the case of structured mixing models. The approach in [49] attempts to impose (16) in (9), but problems with the non-orthogonality of the SOI and background

subspaces must be solved by heuristic changes to the gradient of the contrast function. Similarly, the derivation in [50] intuitively uses (16); however, the way the original blind algorithm is modified cannot be applied to other algorithms, such as the one proposed in [51]. In contrast, the derivation of iFastIVE presented here offers a cleaner interpretation compared to [49], [50], and in the following sections, we will demonstrate that it can be generalized to other structured mixing models.

C. INFORMED PHASE-SHIFT IVE

The contrast function (23) for the semi-blind estimation of $\lambda_1, \ldots, \lambda_K$ was obtained by composing the contrast function (22) and the structured mixing model (6). It is, therefore, convenient to use the complex-valued chain rule [59] for computing the first and second derivatives. First, we note that by (6),

$$\frac{\partial \mathbf{a}_k(\lambda_k)}{\partial \lambda_\ell} = \left(\frac{\partial \mathbf{a}_k^*(\lambda_k)}{\partial \lambda_\ell}\right)^* = \begin{cases} \mathrm{i}(\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k) & \ell = k \\ \mathbf{0} & \ell \neq k \end{cases}, \tag{43}$$

where $\odot$ denotes the Hadamard (element-wise) product. Using this and assuming that the conditions of Statement 1 are met, we can use (26), and by the chain rule we have that

$$\frac{\partial \mathcal{C}_{\boldsymbol{\lambda}}}{\partial \lambda_k} = \sum_{\ell=1}^{K} \left(\frac{\partial \mathcal{C}_{\mathbf{a}}}{\partial \mathbf{a}_\ell}\right)^T \frac{\partial \mathbf{a}_\ell(\lambda_\ell)}{\partial \lambda_k} + \left(\frac{\partial \mathcal{C}_{\mathbf{a}}}{\partial \mathbf{a}_\ell^*}\right)^T \frac{\partial \mathbf{a}_\ell^*(\lambda_\ell)}{\partial \lambda_k} \tag{44}$$

$$= \boldsymbol{\Delta}_k^H \frac{\partial \mathbf{a}_k(\lambda_k)}{\partial \lambda_k} + \boldsymbol{\Delta}_k^T \frac{\partial \mathbf{a}_k^*(\lambda_k)}{\partial \lambda_k} \tag{45}$$

$$= -2\Im\left\{\boldsymbol{\Delta}_k^H (\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k)\right\}, \tag{46}$$

where $\Im\{\cdot\}$ denotes the imaginary part of the argument.

For obtaining a second-order derivative of $\mathcal{C}_{\boldsymbol{\lambda}}$, we want to use the results of Statement 2. Therefore, we will consider the value of the derivative when $N \to +\infty$ and $\lambda_k = \dot{\lambda}_k$. Thus,

$$\frac{\partial^2 \mathcal{C}_{\boldsymbol{\lambda}}}{\partial \lambda_k^2} = -2\Im\left\{\frac{\partial \boldsymbol{\Delta}_k^H}{\partial \lambda_k}(\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k) + \mathrm{i}\boldsymbol{\Delta}_k^H(\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k^2)\right\},$$

$$= -2\Im\left\{\frac{\partial \boldsymbol{\Delta}_k^H}{\partial \lambda_k}(\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k)\right\}, \tag{47}$$

where $\mathbf{v}^2 = \mathbf{v} \odot \mathbf{v}$. Note that the latter equation in (47) holds because $\boldsymbol{\Delta}_k = \mathbf{0}$ for $N \to +\infty$ and $\lambda_k = \dot{\lambda}_k$. To express $\frac{\partial \boldsymbol{\Delta}_k}{\partial \lambda_k}$, we apply the chain rule and use (34)-(36) and (43), which gives

$$\frac{\partial \boldsymbol{\Delta}_k}{\partial \lambda_k} = \sum_{\ell=1}^{K} \left(\frac{\partial \boldsymbol{\Delta}_k^T}{\partial \mathbf{a}_\ell}\right)^T \frac{\partial \mathbf{a}_\ell(\lambda_\ell)}{\lambda_k} + \left(\frac{\partial \boldsymbol{\Delta}_k^T}{\partial \mathbf{a}_\ell^*}\right)^T \frac{\partial \mathbf{a}_\ell^*(\lambda_\ell)}{\lambda_k} \tag{48}$$

$$= \mathbf{H}_k^T \frac{\partial \mathbf{a}_k(\lambda_k)}{\lambda_k} = \mathrm{i}\mathbf{H}_k^*(\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k). \tag{49}$$

By putting (49) into (47),

$$\frac{\partial^2 \mathcal{C}_{\boldsymbol{\lambda}}}{\partial \lambda_k^2} = -2(\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k)^H \mathbf{H}_k^*(\mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k), \tag{50}$$

where the imaginary part operator $\Im\{\cdot\}$ could have been removed since $\mathbf{H}_k^*$ is hermitian.

Note also that from (35)-(36) it follows that the mixed second-order derivatives are zero, i.e. $\frac{\partial^2 \mathcal{C}_{\boldsymbol{\lambda}}}{\partial \lambda_k \partial \lambda_\ell} = 0$ for $k \neq \ell$. Therefore, the Newton-Raphson updates for $\lambda_1, \ldots, \lambda_K$ are separated and can be performed through

$$\lambda_k \leftarrow \lambda_k - \frac{\partial \mathcal{C}_{\boldsymbol{\lambda}}}{\partial \lambda_k /} \frac{\partial^2 \mathcal{C}_{\boldsymbol{\lambda}}}{\partial \lambda_k^2}, \quad k = 1, \ldots, K, \tag{51}$$

where the derivatives are given by (46) and (50). Similarly to the previous algorithm, we have to cope with the unknown values in $\mathbf{H}_k$ that appear in (50). Here again, we replace them with their current estimates. In addition, we found it necessary to replace the value of $\dot{\mathbf{w}}_k$ so that the rank of the resulting matrix is $d-1$ (similarly to the rank of the true $\mathbf{H}_k$). To this end, we first rewrite $\mathbf{H}_k$ as

$$\mathbf{H}_k = \frac{\dot{\nu}_k - \dot{\rho}_k}{\dot{\nu}_k} \dot{\sigma}_{\alpha,k}^4 \dot{\mathbf{C}}_{\alpha_k}^{-*} \left(\frac{1}{\dot{\sigma}_k^2} \dot{\mathbf{C}}_{\mathbf{x}_k}^* - \dot{\mathbf{a}}_k^* \dot{\mathbf{a}}_k^T\right) \dot{\mathbf{C}}_{\alpha_k}^{-*}, \tag{52}$$

which we are allowed to do by Lemma 1. Based on this, we replace $\mathbf{H}_k$ by

$$\breve{\mathbf{H}}_k = \frac{\nu_k - \rho_k}{\nu_k} \sigma_{\alpha,k}^4 \mathbf{C}_{\alpha_k}^{-*} \left(\frac{1}{\varsigma_k^2} \mathbf{C}_{\mathbf{x}_k}^* - \mathbf{a}_{\alpha\mathrm{OC},k}^* \mathbf{a}_{\alpha\mathrm{OC},k}^T\right) \mathbf{C}_{\alpha_k}^{-*}, \tag{53}$$

because $\varsigma_k^{-2}\mathbf{C}_{\mathbf{x}_k}^* - \mathbf{a}_{\alpha\mathrm{OC},k}^* \mathbf{a}_{\alpha\mathrm{OC},k}^T$ has rank $d-1$; for proof, see Appendix D. An iterative semi-blind algorithm for the structured model (6), using the constraints (17)-(18) and the NR update rule (51) is proposed in a similar way to iFastIVE. We will refer to it as iPSIVE (Informed Phase-Shift IVE); the pseudo-code is summarized in Algorithm 2.

D. INFORMED FAR-FIELD IVE

Under similar assumptions and using the chain rule, we now easily derive an algorithm for the one-parameter mixing model (8), whose contrast function is given by (24). By (44) and (46), it easily follows that

$$\frac{\partial \mathcal{C}_\lambda}{\partial \lambda} = -2\sum_{k=1}^{K} \Im\left\{\boldsymbol{\Delta}_k^H (\mathbf{a}_k(\lambda) \odot \mathbf{v}_k)\right\}. \tag{54}$$

Similarly to (47), the derivative of (54) reads

$$\frac{\partial^2 \mathcal{C}_\lambda}{\partial \lambda^2} = -2\sum_{k=1}^{K} \Im\left\{\frac{\partial \boldsymbol{\Delta}_k^H}{\partial \lambda}(\mathbf{a}_k(\lambda) \odot \mathbf{v}_k)\right\}, \tag{55}$$

and following (49),

$$\frac{\partial \boldsymbol{\Delta}_k}{\partial \lambda} = \mathbf{H}_k^T \frac{\partial \mathbf{a}_k(\lambda)}{\lambda} = \mathrm{i}\mathbf{H}_k^*(\mathbf{a}_k(\lambda) \odot \mathbf{v}_k). \tag{56}$$

Then, by putting (56) into (55), the second derivative of $\mathcal{C}_\lambda(\lambda)$ based on Statement 2 is

$$\frac{\partial^2 \mathcal{C}_\lambda}{\partial \lambda^2} = -2\sum_{k=1}^{K} (\mathbf{a}_k(\lambda) \odot \mathbf{v}_k)^H \mathbf{H}_k^*(\mathbf{a}_k(\lambda) \odot \mathbf{v}_k), \tag{57}$$

where the unknown exact value of $\mathbf{H}_k$ can be replaced by $\breve{\mathbf{H}}_k$ given by (53). The corresponding algorithm will be referred

Algorithm 2: The iPSIVE Algorithm.

Input: $\mathbf{x}_k(n)$, $\lambda_k = \lambda_k^{\text{ini}}$, $\mathbf{a}_k = \mathbf{a}_k(\lambda_k)$, $\alpha_k(n)$, $\mathbf{v}_k$, $\texttt{tol}$, $n = 1, \ldots, N$
Output: λ_k, $\mathbf{w}_k$, $s_k(n)$

repeat
  for $k = 1, \ldots, K$ **do**
    $\mathbf{a}_{k,\text{old}} = \mathbf{a}_k$;
    $\sigma^2_{\alpha,k} = (\mathbf{a}_k^H \mathbf{C}^{-1}_{\alpha_k} \mathbf{a}_k)^{-1}$;
    $\mathbf{w}_k = \sigma^2_{\alpha,k} \mathbf{C}^{-1}_{\alpha_k} \mathbf{a}_k$;
    $\varsigma_k^2 = \mathbf{w}_k^H \mathbf{C}_{\mathbf{x}_k} \mathbf{w}_k$;
    $\mathbf{a}_k = \varsigma_k^{-2} \mathbf{C}_{\mathbf{x}_k} \mathbf{w}_k$;
    $s_k(n) = \mathbf{w}_k^H \mathbf{x}_k(n)$;
  end
  $\overline{\mathbf{s}}(n) = [\frac{s_1(n)}{\varsigma_1}, \ldots, \frac{s_K(n)}{\varsigma_K}]^T$;
  for $k = 1, \ldots, K$ **do**
    $\nu_k = \mathrm{E}\left[\phi_k\left(\overline{\mathbf{s}}\right) \frac{s_k}{\varsigma_k}\right]$;
    $\rho_k = \mathrm{E}\left[\frac{\partial \phi_k(\overline{\mathbf{s}})}{\partial s_k^*}\right]$;
    $\mathbf{\Delta}_k = \sigma^2_{\alpha,k} \mathbf{C}^{-1}_{\alpha_k} \left(\mathbf{a}_k - \nu_k^{-1} \mathrm{E}\left[\phi_k(\overline{\mathbf{s}}) \cdot \frac{\mathbf{x}_k}{\sigma_k}\right]\right)$;
    $\breve{\mathbf{H}}_k = \frac{\nu_k - \rho_k}{\nu_k} \sigma^4_{\alpha,k} \mathbf{C}^{-*}_{\alpha_k} \left(\frac{1}{\varsigma_k^2} \mathbf{C}^*_{\mathbf{x}_k} - \mathbf{a}_k^* \mathbf{a}_k^T\right) \mathbf{C}^{-*}_{\alpha_k}$;
    $\mathbf{a}_{\mathbf{v},k} = \mathbf{a}_k(\lambda_k) \odot \mathbf{v}_k$;
    $\lambda_k = \lambda_k - \frac{\Im\{\mathbf{\Delta}_k^H \mathbf{a}_{\mathbf{v},k}\}}{\mathbf{a}_{\mathbf{v},k}^H \breve{\mathbf{H}}_k^* \mathbf{a}_{\mathbf{v},k}}$;
    $\mathbf{a}_k = \mathbf{a}_k(\lambda_k)$;
  end
  $\text{crit} = \max_{k=1,\ldots,K} \left\{1 - \frac{|\mathbf{a}_k^H \mathbf{a}_{k,\text{old}}|}{\|\mathbf{a}_k\| \|\mathbf{a}_{k,\text{old}}\|}\right\}$;
until $\text{crit} < \texttt{tol}$;

Algorithm 3: The iCaponIVE Algorithm.

Input: $\mathbf{x}_k(n)$, $\lambda = \lambda^{\text{ini}}$, $\mathbf{a}_k = \mathbf{a}_k(\lambda)$, $\alpha_k(n)$, $\mathbf{v}_k$, $\texttt{tol}$, $n = 1, \ldots, N$
Output: λ, $\mathbf{w}_k$, $s_k(n)$

repeat
  for $k = 1, \ldots, K$ **do**
    lines 3-8 in Algorithm 2;
  end
  $\overline{\mathbf{s}}(n) = [\frac{s_1(n)}{\varsigma_1}, \ldots, \frac{s_K(n)}{\varsigma_K}]^T$;
  for $k = 1, \ldots, K$ **do**
    lines 12-15 in Algorithm 2;
    $\mathbf{a}_{\mathbf{v},k} = \mathbf{a}_k(\lambda) \odot \mathbf{v}_k$;
  end
  $\lambda = \lambda - \frac{\sum_{k=1}^K \Im\{\mathbf{\Delta}_k^H \mathbf{a}_{\mathbf{v},k}\}}{\sum_{k=1}^K \mathbf{a}_{\mathbf{v},k}^H \breve{\mathbf{H}}_k^* \mathbf{a}_{\mathbf{v},k}}$;
  $\mathbf{a}_k = \mathbf{a}_k(\lambda)$;
  $\text{crit} = \max_{k=1,\ldots,K} \left\{1 - \frac{|\mathbf{a}_k^H \mathbf{a}_{k,\text{old}}|}{\|\mathbf{a}_k\| \|\mathbf{a}_{k,\text{old}}\|}\right\}$;
until $\text{crit} < \texttt{tol}$;

to as iCaponIVE, and its pseudo-code is briefly described in Algorithm 3.

V. NUMERICAL VALIDATION

A. INFORMED ALGORITHMS FOR UNSTRUCTURED MIXING MODEL

We here verify the functionality of iFastIVE through the Monte Carlo simulation similar to that described in Section 5.1.1 in [50]. In a trial, mixtures of signals and references are generated at random, and the SOI is extracted by the compared algorithms. The evaluation criteria are the success rate and the signal-to-interference ratio (SIR). The success rate is equal to the percentage of trials where the SIR above 3 dB is achieved; the SIR is averaged over the successful trials. Each setup is repeated in 5000 trials.

The signal mixtures are generated as follows. $K = 6$ complex-valued mixtures each of $d = 6$ independent signals of length $N = 200$ samples are generated and mixed by random mixing matrices. The SOI is drawn from the circular Generalized Gaussian distribution [60] with the shape parameter $\xi = 0.4$ and variance $\sigma_\ell^2 = \sin(\frac{\ell\pi}{L+1})^\tau$ on the lth interval, $\ell = 1, \ldots, L$, $L = 10$; the length of intervals is N/L; $\tau = 2$. The SOIs of the K mixtures are mixed by a random $K \times K$ unitary matrix to make them mutually dependent. The other interference signals are independently drawn from the Laplacean distribution with random variance on the intervals between $\sqrt{0.1}$ and 10. Note that each interfering source can be confused with the target source. The algorithms are initialized in a randomly perturbed true mixing vector with perturbation variance $\delta^2 = 0.1$.

The reference information $r_k(n)$ consists of a noisy variance profile of the SOI given by

$$r_k(n) = \sqrt{1 - \epsilon^2}\tilde{\sigma}_\ell^2 + \epsilon u_k(\ell), \tag{58}$$

where $u_k(\ell) \sim \mathcal{U}(0, 1)$ (the uniform distribution on [0,1]), $\ell = \lceil \frac{nL}{N} \rceil$ is the interval index, and $\tilde{\sigma}_\ell^2$ is the sample variance of $\tilde{s}_k(n) = \sqrt{1 - \epsilon^2} s_k(n) + \epsilon w_k(n)$ on the ℓth interval where $w_k(n) \sim \mathcal{CN}(0, 1)$. The parameter $\epsilon^2 \in [0, 1]$ controls the quality of the side information. The weighting functions are chosen according to $\alpha_k(n) = \frac{1}{\lambda + |r_k(n)|^2}$ with $\lambda = 10^{-3}$.

The results of the simulation, depending on the parameter ϵ^2, are shown in Fig. 1. We compare the blind algorithm FastIVA [13] (FIVA), its informed variant from [50] (iFIVA), the informed variant iFastIVE derived in this article, and the informed auxiliary function-based algorithm [61] (p-AuxIVA).

While the performance of the blind FIVA is independent of ϵ^2, the performances of the other algorithms are best for $\epsilon^2 = 0$, when $r_k(n)$ contains accurate information about the SOI variance, and weakest for $\epsilon^2 = 1$, when $r_k(n)$ contains only noise. For $\epsilon^2 = 0$, the informed algorithms converge globally to the SOI in almost 100% of trials and achieve a higher SIR than the blind FIVA. As ϵ^2 increases, their success rate and SIR decrease, and near $\epsilon^2 = 1$, the results are even worse than those of FIVA.

Fig. 2 shows histograms of the number of iterations required by the algorithms to reach the stopping criterion or the limit of 100 iterations. As the value of ϵ^2 increases, the number of iterations required for convergence generally grows, with the exception of FICA, whose convergence is

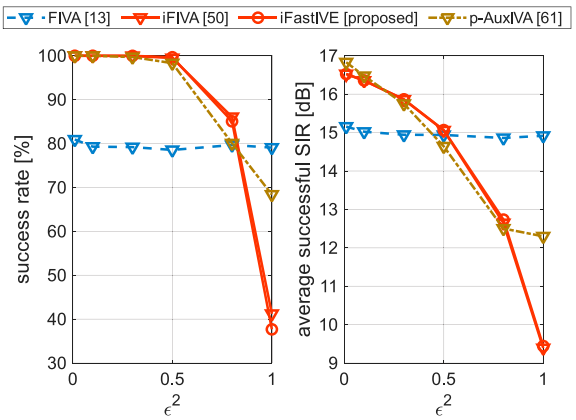

FIGURE 1. Success rate and average SIR of compared algorithms as functions of the quality of prior information controlled through the parameter ϵ^2 (smaller values mean higher quality).

independent of ϵ^2. The proposed iFastIVE and iFIVA exhibit fast convergence within 10 iterations in most trials, unless the prior information is poor ($\epsilon^2 > 0.5$). Such rapid convergence motivates hybridization of algorithms through unrolling, as convergence could be achieved by passing through a small number of layers (see Section VI).

The experiment thus verified that the informed algorithms are capable of using reference information. At the same time, we verified that iFastIVE is comparable to similar algorithms such as iFIVA and p-AuxIVA. This confirms the relevance of the contrast function (22) using the constraint (17)-(18).

B. ALGORITHMS FOR PHASE-SHIFT AND FAR-FIELD MIXING MODELS

We perform a similar simulation to the previous one, with the difference that mixing matrices are generated such that their first column obeys the structure given by (6); we consider $\mathbf{v}_k = [0, 1, 2, \ldots, d-1]^T$ and the ground-truth value $\lambda_k = 0.5, k = 1, \ldots, K$. The data thus obey the phase-shift as well as the far-field mixing models. In this experiment, $K = 5$, $d = 4$, and the reference signals are equal to

$$r_k(n) = \sqrt{1-\epsilon^2}s_k(n) + \epsilon w_k(n), \tag{59}$$

where $w_k(n) \sim \mathcal{CN}(0, 1)$. For simplicity, we select $L = 1$ so all the signals are stationary.

The compared algorithms are the proposed informed algorithms iFastIVE, iPSIVE, and iCaponIVE. We also evaluate their blind variants, i.e., without reference information (as if $r_k(n) = 1$), denoted as FastIVE, PSIVE, and CaponIVE, respectively. The value of ϵ^2 is put equal to 0.4. We investigate algorithms' performances as the length of data N varies from 10 through 1000.

The results in Fig. 3 show that iPSIVE and iCaponIVE, as well as their blind variants, outperform iFastIVE resp. FastIVE when the data is very short (say, $N \leq 50$). For $N > 50$, iCaponIVE shows significantly limited performance compared to the other methods.[7] While iPSIVE still outperforms iFastIVE in terms of SIR as $N > 50$, its success rate grows more slowly with N than that of iFastIVE.

In conclusion, the results suggest that knowledge of the structure of the mixing vector, and thus its reduced parameterization, can be particularly beneficial when little data is available.[8] Another important fact is that the simulations verified the generality of the procedure for deriving iIVE algorithms for structured mixing models. All informed algorithms in the experiment achieve improved results than their blind counterparts.

VI. HYBRID TSE ARCHITECTURES

This section addresses a pilot study in which iFastIVE is incorporated into hybrid trainable architectures. There are many ways to integrate the algorithm, and it depends greatly on the application. Here, we will focus on the basic task of mining side information for iIVE, and the target application is TSE in the time-frequency domain.

A. SCENARIO

The target speaker (SOI) and an interfering speaker are talking simultaneously in a simulated room of dimensions $5 \times 6 \times 2.5$ m. The SOI is located in front of a microphone array of 3 microphones at a random distance in $\langle 0.125, 1.25 \rangle$ m, within the angle $\langle -30, 30 \rangle$ degrees from the axis of the array. The interfering speaker can be located anywhere else in the room, excluding positions directly behind the SOI and behind the array; its minimum distance to the array is 1.5 m. The situation is simulated using the room impulse response generator [63], such that the reverberation time is $T_{60} = 180$ ms. The dry utterances are taken from the development part of the Wall Street Journal dataset (WSJ0-2mix). The reverberated utterances are summed together at the Signal-to-Noise Ratio (SNR) in the range $\langle 2, 10 \rangle$ dB. In this way, 1000 training, 200 validation, and 300 test mixtures of length 5 s are generated.

B. SIDE INFORMATION PROVIDED THROUGH A TRAINED NOISE-ONLY ACTIVITY DETECTOR (NAD)

Intuitively, the weight function $\alpha(n)$ in (15) should emphasize only those moments when only noise and interference are active. This would bring (15) closer to the noise covariance matrix, ensuring that (18) quickly approximates the true separating vector. Therefore, we consider a trained NAD that is directly used as the value of $\alpha(n)$. The NAD is a real-valued deep neural network detecting frames, where the interfering

[7] The results suggest that CaponIVE is not statistically efficient. We do not know the theoretical reason for this, but we conjecture that this might be caused by the mixing parameters being too tightly bound by the orthogonal constraint. Although this reduces the number of parameters to a single real parameter, it is possible that a weaker binding would allow statistical efficiency to be achieved. For example, the performance analysis in [62] shows how orthogonal constraint deteriorates the accuracy of the Symmetric FastICA algorithm.

[8] In the future, this property might be interesting for online adaptive methods, which use the shortest possible data context due to short latency.

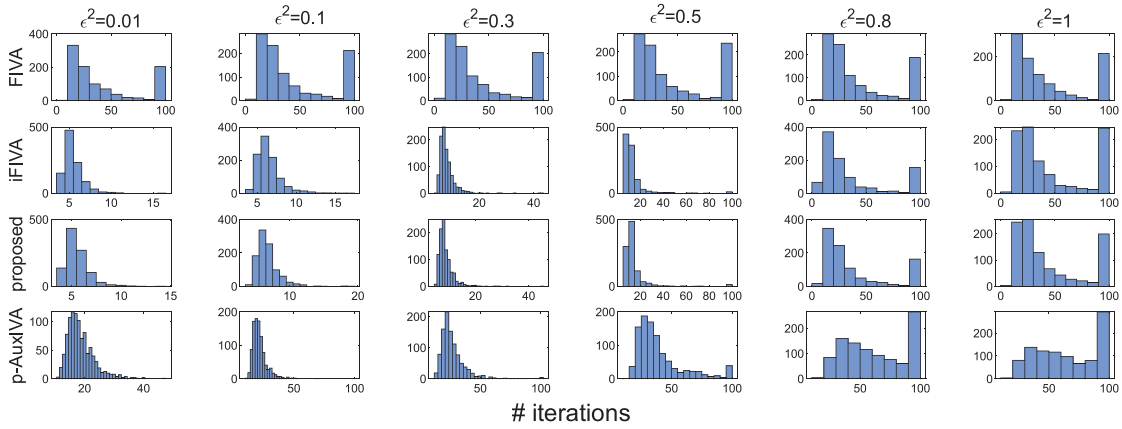

FIGURE 2. Histograms of the number of iterations until the stopping criterion is reached in 1000 trials; the maximum number of iterations is 100. The number of iterations required for convergence generally increases with ϵ, which controls the quality of prior information. A high number of stops at 100 iterations means that the algorithm terminated after reaching the maximum number of iterations, which typically indicates slow convergence or divergence.

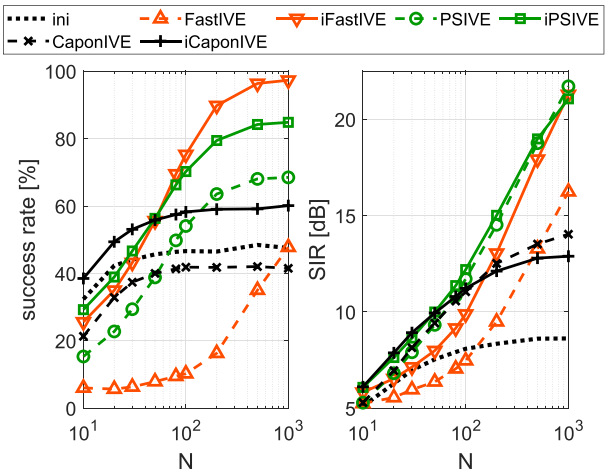

FIGURE 3. Success rate and average SIR of compared algorithms as functions of data length N; the experiment covers extremely short data $N = 10, \ldots, 50$ where structured model-based iPSIVE and iCaponIVE show advantages compared to the generic model-based iFastIVE.

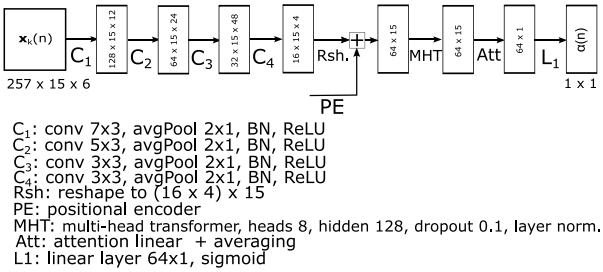

FIGURE 4. NAD network architecture: Each arrow represents a layer or an operation, each rectangle indicates the resulting dimensions of the data (frequency resolution × number of frames × number of feature maps).

source is dominant. Its target value is 1 for frames where the energy of the interferer is by 10 dB greater than that of the SOI, and 0 elsewhere.

The NAD input consists of 3 complex-valued STFT coefficients of the signals from microphones; a noncausal context of 15 frames is used, with 7 frames preceding and 7 following the current frame. The real and imaginary values are split, forming the $2 \times 3 = 6$ input channels. Note that according to the assumed scenario, the SOI and the interferer should be identifiable from the multichannel input data based on their positions.

The NAD architecture is shown in Fig. 4. The processing starts with 4 blocks based on convolutional layers. In each block, a 2-D convolutional layer is followed by an average pooling, reducing the frequency resolution K by a factor of 2; the time-resolution remains unchanged. Subsequently, a batch normalization is applied, followed by the application of the ReLU nonlinearity. Next, learnable positional parameters are added to the 15 frequency-reduced feature vectors to provide explicit temporal information. The features are processed by a Transformer encoder layer with multi-head self-attention, featuring 8 attention heads operating on a 64-dimensional embedding space. The multi-head attention outputs are then fed to a linear attention projection layer that computes attention scores, which are normalized using softmax to produce attention weights. The final encoded feature vector is obtained through a weighted sum of the attention outputs across the temporal dimension. The output of the network is computed as a linear combination of the feature vector elements followed by a sigmoid, which limits the value of the output weighting $\alpha(n)$ to the interval $\langle 0, 1 \rangle$. The NAD network features 55.9 k trainable parameters. Its training proceeds through minimization of the mean square error.

C. FINETUNING THROUGH AN UNROLLED ALGORITHM

Even the target NAD value does not guarantee the optimal choice of the covariance matrix (15). Here, we will therefore try to improve the NAD estimate by directly optimizing the

TABLE 1. Experimental Results for the Baseline Scenario and All Its Variants Concerning Unseen Conditions

	1			2			3						4			5		
	Baseline scenario			Different room dimensions			Reverberation time						Double frequency resolution			Low input SIR		
							$T_{60} = 300$ ms			$T_{60} = 600$ ms								
Method	SDR [dB]	SIR [dB]	STOI [-]	SDR [dB]	SIR [dB]	STOI [-]	SDR [dB]	SIR [dB]	STOI [-]	SDR [dB]	SIR [dB]	STOI [-]	SDR [dB]	SIR [dB]	STOI [-]	SDR [dB]	SIR [dB]	STOI [-]
Mixture	6.0	6.0	0.71	6.0	6.0	0.71	6.0	6.0	0.70	6.0	6.0	0.69	6.0	6.0	0.71	-2.8	-2.8	0.51
FastIVE	-3.4	0.4	0.25	-1.8	2.1	0.29	-3.2	2.1	0.26	-1.8	4.4	0.38	-6.0	0.0	0.11	-3.9	-0.6	0.31
iFastIVE	13.2	17.7	0.79	10.3	15.1	0.73	9.4	14.9	0.72	5.4	11.5	0.62	10.0	18.7	0.64	5.2	7.8	0.63
uFastIVE	13.6	18.0	0.80	10.6	16.5	0.76	10.0	15.2	0.74	6.1	11.9	0.65	9.5	17.8	0.63	5.0	7.4	0.63
iFastIVE matched	•	•	•	12.0	16.9	0.78	9.8	15.4	0.74	6.0	12.4	0.64	9.8	18.3	0.64	6.3	9.5	0.65
uFastIVE matched	•	•	•	12.8	17.4	0.81	10.3	15.5	0.75	7.0	12.6	0.68	11.1	19.3	0.70	6.2	9.7	0.63
oFastIVE	15.2	19.3	0.85	13.9	18.8	0.84	10.6	16.3	0.77	6.8	13.4	0.68	14.0	21.5	0.80	7.3	10.5	0.67
oMWF	14.8	17.8	0.86	14.1	16.7	0.85	11.9	15.1	0.80	10.2	12.5	0.76	13.7	18.5	0.86	10.2	13.2	0.75
ConvTasNet	12.5	21.3	0.83	10.2	17.4	0.78	10.3	17.9	0.77	7.0	13.9	0.66	12.5	21.3	0.83	1.1	5.1	0.58
SpeakerBeam	10.8	17.3	0.80	9.9	16.3	0.78	9.5	16.0	0.75	6.9	13.4	0.65	10.8	17.3	0.80	0.8	4.4	0.57

iFastIVE output. We use the unrolling procedure [64] to create a single neural structure.

The unrolling proceeds as follows. The NAD output is evaluated for the entire context of data $n = 1, \ldots, N$. Then, the result is fed to 5 iterations of iFastIVE. The extraction performance is evaluated using the mean-square error (MSE) loss between the estimated and true SOI. Using backpropagation, we finetune the NAD starting from its pretrained state. The unrolled architecture is referred to as uFastIVE.

D. COMPARED METHODS

In the following experiments, we evaluate the performance of the blind variant of iFastIVE (FastIVE), iFastIVE informed by the output of the pre-trained NAD, and the finetuned uFastIVE. For comparison, two oracle approaches utilizing ground-truth source signals are considered: 1) iFastIVE informed by the ideal target of the NAD as defined in Section VI-B (oFastIVE), and 2) the oracle Multi-channel Wiener filter (oMWF, [65]) corresponding to the optimal spatial filter in the least square sense.

Another important comparison is made with two well-known fully data-driven methods: A) SpeakerBeam (6.7 M parameters, [31]) is a single-channel time-domain target speech extraction model. It identifies the SOI through reference utterances called enrollment. The enrollments originate from WSJ0 sentences that are not included in our train/test sets. B) ConvTasNet (5.0 M parameters, [30]) is a single-channel time-domain separation/enhancement approach recognizing the SOI implicitly through the clean/mixture signal pairs contained in the training data. In the case of our training set, ConvTasNet is likely to focus most on the loudest or least reverberated speaker.

The experiments are evaluated using the BSS_EVAL toolbox [66]. The presented measures are SIR (note that the SIR in BSS_EVAL differs from the criterion used in Section V), which quantifies suppression of the unwanted sources, and SDR, which measures both the suppression and the distortion of the desired source. In addition, the *extended* STOI metric [67] is also evaluated, which measures the short-time objective intelligibility. The given values represent averaged metrics over all 300 test extraction experiments.

E. BASELINE RESULTS

The results for the baseline scenario described in Section VI-A averaged over 300 test mixtures are summarized in the Column 1 of Table 1 (denoted by Table 1-1). They show that iFastIVE and uFastIVE are able to consistently extract the SOI while the blind FastIVE yields poor performance. This is mostly due to the ambiguity of the target speaker, causing the blind method to focus on the interfering speaker in many trials.

The iFastIVE and uFastIVE yield performances approaching that of the oracle methods oFastIVE and oMWF. Note that oFastIVE (i.e., extraction endowed with accurate side information) achieves comparable results to oMWF, which points to the efficiency of iIVE when accurate side information is provided. Although the room for improvement is not large, uFastIVE improves the results compared to iFastIVE by 0.4 dB SDR and 0.01 STOI, demonstrating the effectiveness of optimization through backpropagation. Note that iFastIVE and uFastIVE are trained to the baseline scenario, so their "matched" variants are not relevant in this case (bullets are marked instead of results).

The fully data-driven SpeakerBeam and ConvTasNet show strong results in terms of SIR, where they even outperform oMWF. The iFastIVE and uFastIVE outperform them in terms of SDR. It is worth noting here that (i/u)FastIVE requires two orders of magnitude fewer trainable parameters (55.9 k) compared to SpeakerBeam and ConvTasNet; at the same time, however, they are slightly advantaged by using all three input channels. The comparison is therefore not entirely objective; nevertheless, it does point to the effectiveness of iIVE in terms of the number of trainable parameters.

Similarly, we evaluate algorithms with structured mixing models from Sections IV-C and IV-D. We compare the purely blind variants PSIVE and CaponIVE, informed iPSIVE and iCaponIVE, where the weighting function is the output of the pre-trained NAD, and the informed algorithms endowed by oracle weightings, denoted as oPSIVE and oCaponIVE, respectively.

By comparing the results in Table 2 with those in Table 1-1, we can see that, in the TSE task considered here, the algorithms with the structured mixing models bring

TABLE 2. Baseline Scenario (Far-Field Models)

Method	SDR [dB]	SIR [dB]	STOI [-]
Mixture	6.0	6.0	0.71
PSIVE	-1.9	0.9	0.40
CaponIVE	-2.7	0.0	0.38
iPSIVE	7.6	14.7	0.70
iCaponIVE	6.4	12.9	0.68
oPSIVE	8.1	17.2	0.75
oCaponIVE	8.6	16.7	0.77

weaker performance than those with the unstructured model. The main reason is that the far-field models cannot capture room reverberation and are, therefore, not very suitable for TSE in echoic environments. Nevertheless, the results show significant improvement in performance compared to blind algorithms when side information is used. This confirms the functionality of the informed algorithms iPSIVE and iCaponIVE.

In conclusion, we should mention that direct NAD retraining through an unrolled iPSIVE or iCaponIVE was unsuccessful due to the high sensitivity of these algorithms to poor data conditionality (here we mean cases where some frequencies feature a very high ratio between the largest and the smallest eigenvalue of the covariance matrix of input signals, such that it can cause numerical problems, especially in single-precision arithmetic). This is particularly caused by the inverse covariance matrices that appear in (26) and (38), which cancel each other out in the case of iFastIVE but not in the case of iPSIVE and iCaponIVE. In the future, it will therefore be necessary to address poor data conditionality for these algorithms, e.g., using principal component analysis. At the same time, it remains an open question under what conditions and in what way structured models can be applied to real-life problems such as TSE, where their alternatives are methods using regularized contrasts, such as [21], [22]. This topic is, however, beyond the scope of this article.

F. RESULTS IN UNSEEN CONDITIONS

We now focus on the methods' results under conditions that differ from the baseline scenario for which they were trained. In each experiment, we change only one of the parameters to observe its effect. All other properties, such as utterances, speakers, and their locations, remain the same. We also state the results of methods retrained for the changed conditions. The tables indicate this by the string "matched" after the method name.

1) DIFFERENT ROOM DIMENSIONS

In the experiment here, we change the room dimensions from the original $5 \times 6 \times 2.5$ m to $3 \times 4 \times 2.5$ m. This changes the character and color of the reverberation, while T_{60} remains the same.

The results are shown in Table 1-2. For oMWF, there is a slight decline in SDR/SIR by about 1 dB and in STOI by 0.01, which means that the new room is a bit more difficult for speaker extraction than the baseline. This is confirmed by all methods that have not been adapted to the new conditions: (i/u)FastIVE as well as ConvTasNet and SpeakerBeam, lose 3-4 dB in SDR compared to oMWF. The retraining of (i/u)FastIVE brings an improvement of about 2 dB in SDR and 0.05 in STOI. Consequently, changing the size of the room has an effect on the methods, but the loss in efficiency is not critical.

2) REVERBERATION TIME

This experiment considers two more reverberant scenarios: instead of the $T_{60} = 180$ ms in the baseline, the cases of $T_{60} \in \{300, 600\}$ ms are considered. The results are shown in Table 1-3.

The metric values are significantly lower compared to the baseline scenario, showing that the reverberation time is an important parameter. For example, the oMWF at $T_{60} = 600$ ms achieves metrics by -4.6 dB SDR and -5.3 dB SIR lower compared to the baseline scenario. We note that the STFT length is still 512 samples as in the baseline scenario, which might be too short compared to the length of the impulse responses; the influence of the increased frequency resolution is studied in the next subsection.

There is a larger gap between the FastIVE variants and oMWF, especially in SDR and STOI; oFastIVE yields considerably lower SDR and STOI compared to oMWF as well. For $T_{60} = 600$ ms, the models matched to $T_{60} = 180$ ms are able to improve only SIR compared to the mixture, while SDR is almost unchanged, and STOI even deteriorated. By contrast, the matched uFastIVE yields improvement by $+1$ dB SDR and $+6.6$ dB SIR, with only STOI unchanged.

The mismatched uFastIVE yields comparable results to SpeakerBeam for $T_{60} = 300$ ms and is slightly outperformed (by 0.8 dB SIR and 1.5 dB SDR) for $T_{60} = 600$ ms. ConvTasNet shows the most robust performance in this experiment.

3) DOUBLE FREQUENCY RESOLUTION

Now, we explore the proposed methods' performance when the STFT frequency resolution is doubled, that is, to 1024. However, for comparison with the trained baseline conditions, we must keep the original resolution 512 of the input data to the NAD architectures. The NAD output is used to weight the 1024-resolution STFT frames, which are then input to iFastIVE. Note that ConvTasNet and SpeakerBeam operate with time-domain signals on their inputs, so their performances coincide with those achieved in the baseline setting.

Comparing the results in Table 1-4 to Table 1-1, all the frequency-domain models achieve lower SDR and STOI values, although the time-domain test data are the same in both experiments. There is also a larger SDR/STOI gap between the trained iFastIVE variants and the oracle methods. By examining the behavior of the algorithms in more depth, we observed a sensitivity to poor data conditionality, which is worse at the higher frequency resolution. A solution offers preprocessing using principal component analysis. However,

this problem is beyond the scope of the article and will be addressed in the future.

The matched uFastIVE yields the best results among the proposed methods, improving by 1.3 dB SDR and 0.06 STOI compared to iFastIVE.

4) LOW INPUT SIR

In the baseline scenario, the input SIR ranges within $\langle 2, 10\rangle$ dB, which might be a property of the training data that ConvTasNet and SpeakerBeam rely too heavily on. In this experiment, we therefore modify the input SIR in test data to the range within $\langle -6, 0\rangle$ dB.

Table 1-5 shows that although the input SDR and SIR are now around -3 dB, the improvements of the metrics for the iFastIVE variants remain comparable with the baseline scenario (about 9 dB SDR and 12 dB SIR). The mismatched iFastIVE variants yield only slightly lower metrics compared to the matched ones, by about 1 dB SDR and 2 dB SIR. Anyway, it can be stated that the proposed methods generalize their performance to this scenario. By contrast, the performance of the fully data-driven methods ConvTasNet and SpeakerBeam deteriorates significantly. It seems that under the considered mismatched conditions these methods are not able to identify the target speaker reliably. The uFastIVE trained for the baseline scenario outperforms ConvTasNet and SpeakerBeam by about 4 dB in SDR and $2-3$ dB in SIR.

VII. CONCLUSION

We have shown how iIVE methods can be intuitively derived using constraints between the mixing and separating vectors that approximate the MVDR, using SOI-dependent reference signals. The methods can be effectively coupled with trainable architectures that generalize well to unseen conditions in our experiments. Fully data-driven methods often achieve better performance in some conditions, but this comes at the cost of a two orders of magnitude higher number of trainable parameters and sensitivity to deteriorated performance due to unseen test conditions.

The potential of hybrid methods based on iIVE is far from exhausted. In the future, we plan to integrate deeper with other trainable modules, e.g., at the input as an encoder or at the output as a post-processor. In this article, we have not yet addressed the integration of structured mixing models in hybrid systems, but this option certainly offers further potential.

APPENDIX

A: PROOF OF LEMMA 1

Since α_k only depends on $\dot{s}_k$ and is independent of $\dot{\mathbf{y}}_k$, for $N \to +\infty$, it holds that

$$\mathbf{C}_{\alpha_k} = \mathrm{E}\left[\alpha_k \mathbf{x}_k \mathbf{x}_k^H\right] = \mathrm{E}\left[\alpha_k |s_k|^2\right] \dot{\mathbf{a}}_k \dot{\mathbf{a}}_k^H + \mathrm{E}[\alpha_k] \mathbf{C}_{\mathbf{y}_k}. \quad (60)$$

Without loss on generality, let us assume that $\mathrm{E}[\alpha_k] = 1$. It holds that $\dot{\sigma}^2_{\alpha,k} = \dot{\mathbf{w}}_k^H \mathbf{C}_{\alpha_k} \dot{\mathbf{w}}_k = (\dot{\mathbf{a}}_k^H \mathbf{C}_{\alpha_k}^{-1} \dot{\mathbf{a}}_k)^{-1} = \mathrm{E}[\alpha_k |s_k|^2]$. Since $\mathbf{C}_{\mathbf{x}_k} = \dot{\sigma}_k^2 \dot{\mathbf{a}}_k \dot{\mathbf{a}}_k^H + \mathbf{C}_{\mathbf{y}_k}$, we can rewrite (60) as

$$\mathbf{C}_{\alpha_k} = \mathbf{C}_{\mathbf{x}_k} + (\dot{\sigma}^2_{\alpha,k} - \dot{\sigma}_k^2) \dot{\mathbf{a}}_k \dot{\mathbf{a}}_k^H.$$

By the Woodbury matrix identity,

$$\mathbf{C}_{\alpha_k}^{-1} = \mathbf{C}_{\mathbf{x}_k}^{-1} - \mathbf{C}_{\mathbf{x}_k}^{-1} \dot{\mathbf{a}}_k \dot{\mathbf{a}}_k^H \mathbf{C}_{\mathbf{x}_k}^{-1} \left(\frac{\dot{\sigma}_k^2 (\dot{\sigma}^2_{\alpha,k} - \dot{\sigma}_k^2)}{\dot{\sigma}^2_{\alpha,k}} \right). \quad (61)$$

Using the fact that $\dot{\sigma}_k^2 \mathbf{C}_{\mathbf{x}_k}^{-1} \dot{\mathbf{a}}_k = \dot{\mathbf{w}}_k$ (see Section III-C1) and that $\dot{\mathbf{w}}_k^H \dot{\mathbf{a}}_k = 1$, by multiplying (61) by $\dot{\sigma}^2_{\alpha,k} \dot{\mathbf{a}}_k$ from right, we obtain

$$\dot{\sigma}^2_{\alpha,k} \mathbf{C}_{\alpha_k}^{-1} \dot{\mathbf{a}}_k = \dot{\sigma}^2_{\alpha,k} \mathbf{C}_{\mathbf{x}_k}^{-1} \dot{\mathbf{a}}_k - \mathbf{C}_{\mathbf{x}_k}^{-1} \dot{\mathbf{a}}_k (\dot{\sigma}^2_{\alpha,k} - \dot{\sigma}_k^2) = \dot{\mathbf{w}}_k, \quad (62)$$

which is the assertion of the lemma. ■

B: PROOF OF STATEMENT 1

We are going to evaluate the first Wirtinger derivative of $\mathcal{C}_{\mathbf{a}_k}$ by $\mathbf{a}_k^*$ under the assumptions of the statement. $\mathcal{C}_{\mathbf{a}_k}$ is a compound of (9) with the constraint $\mathbf{a}_k \leftarrow \mathbf{a}_{\alpha\mathrm{OC},k}(\mathbf{a}_k)$ and $\mathbf{w}_k \leftarrow \mathbf{w}_{\alpha,k}(\mathbf{a}_k)$ defined by (17)-(18). The expression subject to the derivative consists of four terms:

$$\mathcal{C}_{\mathbf{a}_k}(\mathbf{a}_1, \ldots, \mathbf{a}_K) = \mathrm{E}\left[\log f(\bar{\mathbf{s}})\right] - \sum_{k=1}^{K} \log \varsigma_k^2 - \sum_{k=1}^{K} \mathrm{E}\left[\mathbf{z}_k^H \dot{\mathbf{C}}_{\mathbf{z}_k}^{-1} \mathbf{z}_k\right] + (d-2) \sum_{k=1}^{K} \log |\gamma_{\alpha\mathrm{OC},k}|^2, \quad (63)$$

where $\bar{\mathbf{s}} = [\frac{s_1}{\varsigma_1}, \ldots, \frac{s_K}{\varsigma_K}]^T$,

$$s_k = \mathbf{w}_{\alpha,k}^H \mathbf{x}_k = \sigma^2_{\alpha,k} \mathbf{a}_k^H \mathbf{C}_{\alpha_k}^{-1} \mathbf{x}_k, \quad (64)$$

$$\mathbf{z}_k = \mathbf{B}(\mathbf{a}_{\alpha\mathrm{OC},k}) \mathbf{x}_k = \frac{\sigma^2_{\alpha,k}}{\varsigma_k^2} \mathbf{B}(\mathbf{R}_k \mathbf{a}_k) \mathbf{x}_k, \quad (65)$$

$$\gamma_{\alpha\mathrm{OC},k} = \mathbf{e}_1^H \mathbf{a}_{\alpha\mathrm{OC},k} = \frac{\sigma^2_{\alpha,k}}{\varsigma_k^2} \mathbf{e}_1^H \mathbf{R}_k \mathbf{a}_k, \quad (66)$$

where $\mathbf{R}_k = \mathbf{C}_{\mathbf{x}_k} \mathbf{C}_{\alpha_k}^{-1}$; $\mathbf{e}_1$ denotes the unit vector (the first column of $\mathbf{I}_d$). We will use the following identities:

$$\frac{\partial}{\partial \mathbf{a}_k^*} \sigma^2_{\alpha,k} = -\sigma^2_{\alpha,k} \mathbf{w}_{\alpha,k}, \quad (67)$$

$$\frac{\partial}{\partial \mathbf{a}_k^*} \varsigma_k = \varsigma_k \left(\frac{1}{2} \frac{\sigma^2_{\alpha,k}}{\varsigma_k^2} \mathbf{C}_{\alpha_k}^{-1} \mathbf{C}_{\mathbf{x}_k} - \mathbf{I}_d \right) \mathbf{w}_{\alpha,k}, \quad (68)$$

$$\frac{\partial}{\partial \mathbf{a}_k^*} s_k = -s_k \mathbf{w}_{\alpha,k} + \sigma^2_{\alpha,k} \mathbf{C}_{\alpha_k}^{-1} \mathbf{x}_k, \quad (69)$$

$$\frac{\partial}{\partial \mathbf{a}_k^*} s_k^* = -s_k^* \mathbf{w}_{\alpha,k}, \quad (70)$$

which the reader can verify from definitions. The derivative of the first term in (63) reads

$$\frac{\partial}{\partial \mathbf{a}_k^*} \mathrm{E}\left[\log f(\bar{\mathbf{s}})\right] = \mathrm{E}\left[\frac{\partial \log f(\bar{\mathbf{s}})}{\partial s_k} \left(\frac{\partial s_k}{\partial \mathbf{a}_k^*} \frac{1}{\varsigma_k} + s_k \frac{\partial \varsigma_k^{-1}}{\partial \mathbf{a}_k^*} \right) + \frac{\partial \log f(\bar{\mathbf{s}})}{\partial s_k^*} \left(\frac{\partial s_k^*}{\partial \mathbf{a}_k^*} \frac{1}{\varsigma_k} + s_k^* \frac{\partial \varsigma_k^{-1}}{\partial \mathbf{a}_k^*} \right) \right]. \quad (71)$$

By the assumption of Statement 1, we apply the replacement $\frac{\partial \log f(\bar{\mathbf{s}})}{\partial s_k} \leftarrow -\nu_k^{-1} \phi_k(\bar{\mathbf{s}})$, thus also $\frac{\partial \log f(\bar{\mathbf{s}})}{\partial s_k^*} \leftarrow -\nu_k^{-*} \phi_k^*(\bar{\mathbf{s}})$. Then, by using the identities (67)-(70) and using the fact that $\nu_k^{-1} \mathrm{E}[\phi_k(\bar{\mathbf{s}}) \frac{s_k}{\varsigma_k}] = 1$, which follows from the definition (25),

after some algebra we obtain

$$\begin{aligned}&\tfrac{\partial}{\partial \mathbf{a}_k^*}\mathrm{E}\left[\log f\left(\bar{\mathbf{s}}\right)\right]\\&=\tfrac{\sigma_{\alpha,k}^2}{\varsigma_k^2}\mathbf{C}_{\alpha_k}^{-1}\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\alpha,k}-\nu_k^{-1}\sigma_{\alpha,k}^2\mathbf{C}_{\alpha_k}^{-1}\mathrm{E}\left[\phi_k(\mathbf{s})\tfrac{\mathbf{x}_k}{\varsigma_k}\right]\\&=\sigma_{\alpha,k}^2\mathbf{C}_{\alpha_k}^{-1}\left(\mathbf{a}_{\alpha\mathrm{OC},k}-\nu_k^{-1}\mathrm{E}\left[\phi_k(\mathbf{s})\tfrac{\mathbf{x}_k}{\varsigma_k}\right]\right).\end{aligned}\tag{72}$$

This result already corresponds with (26), however, there are still terms 2-4 in (63) to be taken into account. In the following, we show that the sum of their derivatives is zero. This requires the orthogonality condition, thus demonstrating the importance of the choice of the constraint (17)-(18).

Terms 2-4 in (63) are separated with respect to k. Therefore, we can omit the index k for now and deal with the case as if $K=1$. As for term 2, using (68), we obtain

$$\frac{\partial}{\partial \mathbf{a}^*}\log\varsigma^2=-2\frac{1}{\varsigma}\frac{\partial\varsigma}{\partial\mathbf{a}^*}=2\mathbf{w}_\alpha-\mathbf{w}_{\alpha\mathrm{OC}},\tag{73}$$

where we use the definition

$$\mathbf{w}_{\alpha\mathrm{OC},k}=\tfrac{\sigma_{\alpha,k}^2}{\varsigma_k^2}\mathbf{C}_{\alpha_k}^{-1}\mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\alpha,k}=\tfrac{\sigma_{\alpha_k}^2}{\varsigma_k^2}\mathbf{R}_k^H\mathbf{w}_{\alpha,k}.\tag{74}$$

Let $\mathbf{E}$ be the matrix such that $\mathbf{I}_d=\begin{bmatrix}\mathbf{e}_1^H\\ \mathbf{E}\end{bmatrix}$ (also, $\mathbf{E}=\begin{bmatrix}\mathbf{0}\ \mathbf{I}_{d-1}\end{bmatrix}$), and let $\tilde{x}=\mathbf{e}_1^H\mathbf{x}$ and $\tilde{\mathbf{x}}=\mathbf{E}\mathbf{x}$, that is, $\mathbf{x}=\begin{bmatrix}\tilde{x}\\ \tilde{\mathbf{x}}\end{bmatrix}$. For the computation of the derivative of the third term of (63), we will use the following definitions and identities:

$$\tilde{\mathbf{z}}=\mathbf{B}(\mathbf{R}\mathbf{a})\mathbf{x}=\tilde{x}\mathbf{E}\mathbf{R}\mathbf{a}-(\mathbf{e}_1^H\mathbf{R}\mathbf{a})\tilde{\mathbf{x}}\tag{75}$$

$$\tfrac{\partial}{\partial\mathbf{a}^*}\tilde{\mathbf{z}}^H=\tilde{x}^*\mathbf{R}^H\mathbf{E}^H-\mathbf{R}^H\mathbf{e}_1\tilde{\mathbf{x}}^H,\tag{76}$$

$$\mathbf{z}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}=\tfrac{\sigma_\alpha^4}{\varsigma^4}\tilde{\mathbf{z}}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\tilde{\mathbf{z}}.\tag{77}$$

Now,

$$\begin{aligned}\tfrac{\partial}{\partial\mathbf{a}^*}\mathrm{E}[\mathbf{z}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}]&=\left(\tfrac{2\sigma_\alpha^2}{\varsigma^4}\tfrac{\partial\sigma_\alpha^2}{\partial\mathbf{a}^*}-4\tfrac{\sigma_\alpha^4}{\varsigma^5}\tfrac{\partial\varsigma}{\partial\mathbf{a}^*}\right)\mathrm{E}[\tilde{\mathbf{z}}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\tilde{\mathbf{z}}]\\&\quad+\tfrac{\sigma_\alpha^4}{\varsigma^4}\tfrac{\partial}{\partial\mathbf{a}^*}\mathrm{E}[\tilde{\mathbf{z}}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\tilde{\mathbf{z}}]\\&=(-2\mathbf{w}_\alpha-2\mathbf{w}_{\alpha\mathrm{OC}}+4\mathbf{w}_\alpha)\mathrm{E}[\mathbf{z}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}]\\&\quad+\tfrac{\sigma_\alpha^2}{\varsigma^2}\left(\mathrm{E}[\tilde{x}^*\mathbf{R}^H\mathbf{E}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}]\right.\\&\quad\left.-\mathrm{E}[\mathbf{R}^H\mathbf{e}_1\tilde{\mathbf{x}}^H\dot{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}]\right).\end{aligned}\tag{78}$$

For further computations, we remind that, according to (3), the observed signals $\mathbf{x}$ can always be written as $\mathbf{x}=\mathbf{a}\begin{bmatrix}s\\ \mathbf{z}\end{bmatrix}$ where $s=\mathbf{w}_\alpha^H\mathbf{x}$, $\mathbf{z}=\mathbf{B}(\mathbf{a}_{\alpha\mathrm{OC}})\mathbf{x}$, and $\mathbf{A}$ is parameterized by the couple of vectors $\mathbf{w}_\alpha$ and $\mathbf{a}_{\alpha\mathrm{OC}}$ according to (4). We now exploit the fact that the subspaces generated by s and $\mathbf{z}$ defined this way are orthogonal (see Section III-D), which means that $\mathrm{E}[\mathbf{z}s^*]=\mathbf{0}$. Therefore,

$$\mathrm{E}[\mathbf{z}\mathbf{x}^H]=\left[\mathrm{E}[\mathbf{z}s^*]\ \mathrm{E}[\mathbf{z}\mathbf{z}^H]\right]\mathbf{A}^H=[\mathbf{0}\ \mathbf{C}_\mathbf{z}]\mathbf{A}^H,\tag{79}$$

where $\mathbf{C}_\mathbf{z}=\mathrm{E}[\mathbf{z}\mathbf{z}^H]$ is the sample-based covariance of the current estimate of $\mathbf{z}$. By the assumption of Statement 1, we replace the unknown $\dot{\mathbf{C}}_\mathbf{z}$ by $\mathbf{C}_\mathbf{z}$. By (79)

$$\mathbf{C}_\mathbf{z}^{-1}\mathrm{E}[\mathbf{z}\mathbf{x}^H]=[\mathbf{0}\ \mathbf{I}_{d-1}]\mathbf{A}^H=\mathbf{E}\mathbf{A}^H.\tag{80}$$

By (4), it holds that $\mathbf{E}\mathbf{A}^H\mathbf{e}_1=\mathbf{E}\mathbf{w}_\alpha=\mathbf{h}_\alpha$ and $\mathbf{E}\mathbf{A}^H\mathbf{E}^H=\gamma_{\alpha\mathrm{OC}}^{-*}(\mathbf{h}_\alpha\mathbf{g}_{\alpha\mathrm{OC}}^H-\mathbf{I}_{d-1})$ where $\mathbf{g}_{\alpha\mathrm{OC}}=\mathbf{E}\mathbf{a}_{\alpha\mathrm{OC}}$. Then,

$$\begin{aligned}\mathrm{E}[\tilde{x}^*\mathbf{R}^H\mathbf{E}^H\mathbf{C}_\mathbf{z}^{-1}\mathbf{z}]&=\mathbf{R}^H\mathbf{E}^H\mathbf{C}_\mathbf{z}^{-1}\mathrm{E}[\mathbf{z}\mathbf{x}^H]\mathbf{e}_1\\&=\mathbf{R}^H\mathbf{E}^H\mathbf{h}_\alpha=\mathbf{R}^H\begin{bmatrix}0\\ \mathbf{h}_\alpha\end{bmatrix},\\\mathrm{E}[\mathbf{R}^H\mathbf{e}_1\tilde{\mathbf{x}}^H\mathbf{C}_\mathbf{z}^{-1}\mathbf{z}]&=\mathbf{R}^H\mathbf{e}_1\mathtt{tr}\left(\mathbf{C}_\mathbf{z}^{-1}\mathrm{E}[\mathbf{z}\mathbf{x}^H]\mathbf{E}^H\right)\qquad(81)\\&=\mathbf{R}^H\mathbf{e}_1\mathtt{tr}\left(\gamma_{\alpha\mathrm{OC}}^{-*}(\mathbf{h}_\alpha\mathbf{g}_{\alpha\mathrm{OC}}^H-\mathbf{I}_{d-1})\right)\\&=\mathbf{R}^H\mathbf{e}_1(-\beta_\alpha-\gamma_{\alpha\mathrm{OC}}^{-*}(d-2)),\qquad(82)\end{aligned}$$

where we have used (5); $\mathtt{tr}(\cdot)$ denotes the trace of the argument. By putting the latter identities into (78) together with the substitution $\dot{\mathbf{C}}_\mathbf{z}\leftarrow\mathbf{C}_\mathbf{z}$ and using that $\mathrm{E}[\mathbf{z}^H\mathbf{C}_\mathbf{z}^{-1}\mathbf{z}]=\mathtt{tr}(\mathbf{C}_\mathbf{z}^{-1}\mathrm{E}[\mathbf{z}\mathbf{z}^H])=d-1$,

$$\begin{aligned}&\tfrac{\partial}{\partial\mathbf{a}^*}\mathrm{E}[\mathbf{z}^H\dot{\mathbf{C}}_\mathbf{z}^{-1}\mathbf{z}]=2(d-1)(\mathbf{w}_\alpha-\mathbf{w}_{\alpha\mathrm{OC}})\\&+\tfrac{\sigma_\alpha^2}{\varsigma^2}\mathbf{R}^H\left(\underbrace{\begin{bmatrix}0\\ \mathbf{h}_\alpha\end{bmatrix}+\beta_\alpha\mathbf{e}_1}_{\mathbf{w}_\alpha}+(d-2)\gamma_{\alpha\mathrm{OC}}^{-*}\mathbf{e}_1\right)\\&=2(d-1)(\mathbf{w}_\alpha-\mathbf{w}_{\alpha\mathrm{OC}})+\mathbf{w}_{\alpha\mathrm{OC}}+\tfrac{d-2}{\gamma_{\alpha\mathrm{OC}}^*}\tfrac{\sigma_\alpha^2}{\varsigma^2}\mathbf{R}^H\mathbf{e}_1.\end{aligned}\tag{83}$$

Finally, we compute the derivative of the last term in (63). Using (66),

$$\begin{aligned}&(d-2)\tfrac{\partial}{\partial\mathbf{a}^*}\log|\gamma_{\alpha\mathrm{OC}}|^2\\&\quad=(d-2)\tfrac{\partial}{\partial\mathbf{a}^*}\left(\log\tfrac{\sigma_\alpha^4}{\varsigma^4}+\log\mathbf{a}^H\mathbf{R}^H\mathbf{e}_1\right).\end{aligned}\tag{84}$$

Using the part of (78), where the derivative of $\frac{\sigma_\alpha^4}{\varsigma^4}$ appears, and using the fact that (66) can be written as $\mathbf{a}^H\mathbf{R}^H\mathbf{e}_1=\frac{\varsigma^2}{\sigma_\alpha^2}\gamma_{\alpha\mathrm{OC}}^*$,

$$\begin{aligned}&(d-2)\tfrac{\partial}{\partial\mathbf{a}^*}\log|\gamma_{\alpha\mathrm{OC}}|^2\\&\quad=2(d-2)(\mathbf{w}_\alpha-\mathbf{w}_{\alpha\mathrm{OC}})+\tfrac{d-2}{\gamma_{\alpha\mathrm{OC}}^*}\tfrac{\sigma_\alpha^2}{\varsigma^2}\mathbf{R}^H\mathbf{e}_1.\end{aligned}\tag{85}$$

By using (73), (83) and (85) in (63), we can see that the derivatives of terms 2-4 in (63) sum to zero, which concludes the proof. ■

C: PROOF OF STATEMENT 2

The assumptions of the statement say that $\mathbf{a}_k=\dot{\mathbf{a}}_k$ and $N\to+\infty$. In this proof, we will, therefore, mostly work with true values of parameters and signals. For simplicity, we will not use the dot accents here, unless confusion can arise.

We start by proving (34).

$$\frac{\partial\boldsymbol{\Delta}_k^T}{\partial\mathbf{a}_k}=\frac{\partial\sigma_\alpha^2}{\partial\mathbf{a}_k}\frac{\boldsymbol{\Delta}_k^T}{\sigma_\alpha^2}+\sigma_{\alpha,k}^2\frac{\partial}{\partial\mathbf{a}_k}\left(\mathbf{a}_{\alpha\mathrm{OC},k}^T\right.$$

$$-\nu_k^{-1}\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k^T}{\varsigma_k}\right]\Bigg)\mathbf{C}_{\alpha_k}^{-*}=\sigma_{\alpha,k}^2\left(\frac{\partial\frac{\sigma_{\alpha,k}^2}{\varsigma_k^2}}{\partial\mathbf{a}_k}\mathbf{a}_k^T\mathbf{R}_k^T\right.$$

$$\left.+\frac{\sigma_{\alpha,k}^2}{\varsigma_k^2}\mathbf{R}_k^T-\nu_k^{-1}\frac{\partial}{\partial\mathbf{a}_k}\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k^T}{\varsigma_k}\right]\right)\mathbf{C}_{\alpha_k}^{-*},\tag{86}$$

since $\boldsymbol{\Delta}_k=\mathbf{0}$ for $\mathbf{a}_k=\dot{\mathbf{a}}_k$ and $N\to+\infty$. In further computations, we use (67)–(70) and derive the following identities:

$$\frac{\partial}{\partial\mathbf{a}_k}\frac{\sigma_{\alpha,k}^2}{\varsigma_k^2}=\frac{\sigma_{\alpha,k}^2}{\varsigma_k^2}(\mathbf{w}_{\alpha,k}^*-\mathbf{w}_{\alpha\mathrm{OC},k}^*),\tag{87}$$

$$\begin{aligned}\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_k}&=\frac{\partial\phi_k}{\partial s_k}\left(\frac{\partial s_k}{\partial\mathbf{a}_k}\frac{1}{\varsigma_k}-\frac{s_k}{\varsigma_k^2}\frac{\partial\varsigma_k}{\partial\mathbf{a}_k}\right)+\frac{\partial\phi_k}{\partial s_k^*}\left(\frac{\partial s_k^*}{\partial\mathbf{a}_k}\frac{1}{\varsigma_k}-\frac{s_k^*}{\varsigma_k^2}\frac{\partial\varsigma_k}{\partial\mathbf{a}_k}\right)\\&=-\frac{1}{2}\left(\frac{\partial\phi_k}{\partial s_k}\frac{s_k}{\varsigma_k}+\frac{\partial\phi_k}{\partial s_k^*}\frac{s_k^*}{\varsigma_k}\right)\mathbf{w}_{\alpha\mathrm{OC},k}^*+\sigma_{\alpha,k}^2\mathbf{C}_{\alpha,k}^{-*}\frac{\mathbf{x}_k^*}{\varsigma_k}\frac{\partial\phi_k}{\partial s_k^*},\end{aligned}\tag{88}$$

$$\frac{\partial}{\partial\mathbf{a}_k}\frac{\mathbf{x}_k^T}{\varsigma_k}=\left(-\tfrac{1}{2}\mathbf{w}_{\alpha\mathrm{OC},k}^*+\mathbf{w}_{\alpha,k}^*\right)\frac{\mathbf{x}_k^T}{\varsigma_k}.\tag{89}$$

For $s_k=\dot{s}_k$ and $\mathbf{z}_k=\dot{\mathbf{z}}_k$, and $\varsigma_k^2=\dot{\varsigma}_k^2=\dot{\sigma}_k^2$, we have the identities that

$$\mathrm{E}\left[\frac{s_k^*}{\varsigma_k}\frac{\mathbf{x}_k^T}{\varsigma_k}\frac{\partial\phi_k}{\partial s_k^*}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\right]=\xi_k\mathbf{a}_k^T,\tag{90}$$

$$\mathrm{E}\left[\frac{s_k}{\varsigma_k}\frac{\mathbf{x}_k^T}{\varsigma_k}\frac{\partial\phi_k}{\partial s_k}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\right]=\eta_k\mathbf{a}_k^T,\tag{91}$$

$$\begin{aligned}\mathrm{E}\left[\frac{\partial\phi_k}{\partial s_k^*}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\frac{\mathbf{x}_k^*}{\varsigma_k}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]&=\xi_k\mathbf{a}_k^*\mathbf{a}_k^T+\rho_k\frac{\mathbf{C}_{\mathbf{y}_k}^*}{\sigma_k^2}\\&=\xi_k\mathbf{a}_k^*\mathbf{a}_k^T+\rho_k\left(\frac{\mathbf{C}_{\mathbf{x}_k}^*}{\sigma_k^2}-\mathbf{a}_k^*\mathbf{a}_k^T\right),\end{aligned}\tag{92}$$

where $\mathbf{C}_{\mathbf{y}_k}=\mathrm{E}[\mathbf{y}_k\mathbf{y}_k^H]$, and we have used the fact that $\mathbf{C}_{\mathbf{x}_k}=\sigma_k^2\mathbf{a}_k\mathbf{a}_k^H+\mathbf{C}_{\mathbf{y}_k}$. Now, we compute the derivative of the expectation value in (86). By using (88)-(92),

$$\begin{aligned}&\frac{\partial}{\partial\mathbf{a}_k}\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k^T}{\varsigma_k}\right]=\mathrm{E}\left[\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_k}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]+\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\frac{\partial}{\partial\mathbf{a}_k}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]=\\&\quad-\tfrac{1}{2}(\eta_k+\xi_k)\mathbf{w}_{\alpha\mathrm{OC},k}^*\mathbf{a}_k^T+\sigma_{\alpha,k}^2\mathbf{C}_{\alpha_k}^{-*}\left(\xi_k\mathbf{a}_k^*\mathbf{a}_k^T\right.\\&\quad\left.+\rho_k\left(\frac{\mathbf{C}_{\mathbf{x}_k}^*}{\sigma_k^2}-\mathbf{a}_k^*\mathbf{a}_k^T\right)\right)+\nu_k\left(-\tfrac{1}{2}\mathbf{w}_{\alpha\mathrm{OC},k}^*\mathbf{a}_k^T+\mathbf{w}_{\alpha,k}^*\mathbf{a}_k^T\right)\\&=-\tfrac{\eta_k+\xi_k+\nu_k}{2}\mathbf{w}_{\alpha\mathrm{OC},k}^*\mathbf{a}_k^T+(\xi_k+\nu_k-\rho_k)\mathbf{w}_{\alpha,k}^*\mathbf{a}_k^T\\&\quad+\rho_k\frac{\sigma_{\alpha,k}^2}{\sigma_k^2}\mathbf{R}_k^T.\end{aligned}\tag{93}$$

By the assumptions of the statement and the assertions of Lemma 1 and 2, $\mathbf{w}_{\alpha,k}=\dot{\mathbf{w}}_k$, $\mathbf{w}_{\alpha\mathrm{OC},k}=\sigma_{\alpha,k}^2\mathbf{C}_{\alpha_k}^{-1}\mathbf{a}_k=\dot{\mathbf{w}}_k$ and $\varsigma_k^2=\sigma_k^2$. Hence, by putting (87) and (93) into (86),

$$\begin{aligned}\frac{\partial\boldsymbol{\Delta}_k^T}{\partial\mathbf{a}_k}&=\sigma_{\alpha,k}^2\left(\frac{\nu_k-\rho_k}{\nu_k}\left(\frac{\sigma_{\alpha,k}^2}{\sigma_k^2}\mathbf{R}_k^T-\mathbf{w}_{\alpha\mathrm{OC},k}^*\mathbf{a}_k^T\right)\right.\\&\quad\left.+\frac{\xi_k-\eta_k-\nu_k}{2\nu_k}\mathbf{w}_{\alpha\mathrm{OC},k}^*\mathbf{a}_k^T\right)\mathbf{C}_{\alpha_k}^{-*}\\&=\frac{\nu_k-\rho_k}{\nu_k}\left(\frac{\sigma_{\alpha,k}^4}{\sigma_k^2}\mathbf{R}_k^T\mathbf{C}_{\alpha_k}^{-*}-\mathbf{w}_k^*\mathbf{w}_k^T\right)+\frac{\xi_k-\eta_k-\nu_k}{2\nu_k}\mathbf{w}_k^*\mathbf{w}_k^T.\end{aligned}\tag{94}$$

By the condition (32), the latter term is zero and (34) follows.

Now, we continue by computing the second Hessian matrix $\frac{\partial\boldsymbol{\Delta}_k^H}{\partial\mathbf{a}_k}=\left[\frac{\partial\boldsymbol{\Delta}_k^T}{\partial\mathbf{a}_k^*}\right]^*$, which can be easily done by using the previous identities. By (86),

$$\frac{\partial\boldsymbol{\Delta}_k^T}{\partial\mathbf{a}_k^*}=\sigma_{\alpha,k}^2\left(\frac{\partial\frac{\sigma_{\alpha,k}^2}{\varsigma_k^2}}{\partial\mathbf{a}_k^*}\mathbf{a}_k^T\mathbf{R}_k^T-\nu_k^{-1}\frac{\partial}{\partial\mathbf{a}_k^*}\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k^T}{\varsigma_k}\right]\right)\mathbf{C}_{\alpha_k}^{-*},\tag{95}$$

where the derivative in the first term is the conjugate value of (87), which is equal to zero as $N\to+\infty$. Hence, only the second term in (95) should be computed, for which we use the following identity:

$$\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_k}=-\frac{1}{2}\left(\frac{\partial\phi_k}{\partial s_k}\frac{s_k}{\varsigma_k}+\frac{\partial\phi_k}{\partial s_k^*}\frac{s_k^*}{\varsigma_k}\right)\mathbf{w}_{\alpha\mathrm{OC},k}+\sigma_{\alpha,k}^2\mathbf{C}_{\alpha,k}^{-*}\frac{\mathbf{x}_k}{\varsigma_k}\frac{\partial\phi_k}{\partial s_k},\tag{96}$$

where we used (67)-(70) and (88). Next, we will use that

$$\mathrm{E}\left[\frac{\partial\phi_k}{\partial s_k}\Big|_{\mathbf{s}=\bar{\mathbf{s}}}\frac{\mathbf{x}_k}{\varsigma_k}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]=\nu_k\mathbf{a}_k\mathbf{a}_k^T+\mathrm{E}\left[\frac{\partial\phi_k}{\partial s_k}\right]\frac{\mathbf{P}_{\mathbf{y}_k}}{\sigma_k^2}=\nu_k\mathbf{a}_k\mathbf{a}_k^T,\tag{97}$$

because $\mathbf{P}_{\mathbf{y}_k}=\mathrm{E}[\mathbf{y}_k\mathbf{y}_k^T]=\mathbf{0}$ is the pseudo-covariance matrix of $\mathbf{y}_k$, which is zero since $\mathbf{y}_k$ are assumed to be circular Gaussian. Using the latter two identities,

$$\begin{aligned}&\frac{\partial}{\partial\mathbf{a}_k^*}\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\cdot\frac{\mathbf{x}_k^T}{\varsigma_k}\right]=\mathrm{E}\left[\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_k^*}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]+\mathrm{E}\left[\phi_k(\bar{\mathbf{s}})\frac{\partial}{\partial\mathbf{a}_k^*}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]\\&\quad=\tfrac{1}{2}(\eta_k-\xi_k)\mathbf{w}_{\alpha\mathrm{OC},k}\mathbf{a}_k^T+\nu_k\left(-\tfrac{1}{2}\mathbf{w}_{\alpha\mathrm{OC},k}\mathbf{a}_k^T+\mathbf{w}_{\alpha,k}\mathbf{a}_k^T\right)\\&\quad=-\tfrac{\eta_k+\xi_k+\nu_k}{2}\mathbf{w}_{\alpha\mathrm{OC},k}\mathbf{a}_k^T+(\eta_k+\nu_k)\mathbf{w}_{\alpha,k}\mathbf{a}_k^T,\end{aligned}\tag{98}$$

which we put into (95) and, for $N\to+\infty$, obtain

$$\frac{\partial\boldsymbol{\Delta}_k^T}{\partial\mathbf{a}_k^*}=\frac{\xi_k-\eta_k-\nu_k}{2\nu_k}\mathbf{w}_k\mathbf{w}_k^T.\tag{99}$$

Now, by (32), (35) follows.

Finally, to compute $\frac{\partial\boldsymbol{\Delta}_k^T}{\partial\mathbf{a}_\ell}$ where $\ell\neq k$, we only need to find the value of $\mathrm{E}\left[\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_\ell}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]$ since all terms but $\phi_k(\bar{\mathbf{s}})$ in $\boldsymbol{\Delta}_k$ are independent of $\mathbf{a}_\ell$. By (88), it readily follows that

$$\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_\ell}=-\frac{1}{2}\left(\frac{\partial\phi_k}{\partial s_\ell}\frac{s_\ell}{\varsigma_\ell}+\frac{\partial\phi_\ell}{\partial s_\ell^*}\frac{s_\ell^*}{\varsigma_\ell}\right)\mathbf{w}_{\alpha\mathrm{OC},\ell}^*+\sigma_{\alpha,\ell}^2\mathbf{C}_{\alpha,\ell}^{-*}\frac{\mathbf{x}_\ell^*}{\varsigma_\ell}\frac{\partial\phi_\ell}{\partial s_\ell^*}.\tag{100}$$

By the assumption of the statistical model, $\mathrm{E}[\mathbf{y}_k\mathbf{y}_\ell^H]=0$. Straightforward computation using that $\mathbf{w}_{\alpha,\ell}=\mathbf{w}_{\alpha\mathrm{OC},\ell}=\mathbf{w}_\ell$ gives

$$\mathrm{E}\left[\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_\ell}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]=\tfrac{1}{2}(\xi_{\ell,k}-\eta_{\ell,k})\mathbf{w}_\ell^*\mathbf{a}_k^T\tag{101}$$

By using (101),

$$\frac{\partial\boldsymbol{\Delta}_k^T}{\partial\mathbf{a}_\ell}=-\nu_k^{-1}\sigma_{\alpha,k}^2\mathrm{E}\left[\frac{\partial\phi_k(\bar{\mathbf{s}})}{\partial\mathbf{a}_\ell}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]\mathbf{C}_{\alpha_k}^{-*}=\tfrac{1}{2}\frac{\eta_{\ell,k}-\xi_{\ell,k}}{\nu_k}\mathbf{w}_\ell^*\mathbf{w}_k^T,\tag{102}$$

and (36) follows from (33). Similar way it can be shown that

$$\mathrm{E}\left[\frac{\partial \phi_k(\bar{\mathbf{s}})}{\partial \mathbf{a}_\ell^*}\frac{\mathbf{x}_k^T}{\varsigma_k}\right] = \tfrac{1}{2}(\eta_{\ell,k} - \xi_{\ell,k})\mathbf{w}_\ell \mathbf{a}_k^T, \tag{103}$$

hence

$$\frac{\partial \mathbf{\Delta}_k^T}{\partial \mathbf{a}_\ell^*} = -\nu_k^{-1}\sigma_{\alpha,k}^2 \mathrm{E}\left[\frac{\partial \phi_k(\bar{\mathbf{s}})}{\partial \mathbf{a}_\ell^*}\frac{\mathbf{x}_k^T}{\varsigma_k}\right]\mathbf{C}_{\alpha_k}^{-*} = \tfrac{1}{2}\frac{\xi_{\ell,k}-\eta_{\ell,k}}{\nu_k}\mathbf{w}_\ell \mathbf{w}_k^T, \tag{104}$$

which concludes the proof. ■

APPENDIX D:

Provided that $\mathbf{a}_k$ and $\mathbf{w}_k$ form a couple of (estimates of) mixing and separating vectors that satisfy the OC introduced in Section III-C1, $\mathbf{x}_k$ can always be written as

$$\mathbf{x}_k = \mathbf{a}_k \underbrace{\mathbf{w}_k^H \mathbf{x}_k}_{s_k} + \underbrace{(\mathbf{I}_d - \mathbf{a}_k \mathbf{w}_k^H)\mathbf{x}_k}_{\mathbf{y}_k} \tag{105}$$

where s_k and $\mathbf{y}_k$ are orthogonal. Therefore, $\mathbf{C}_{\mathbf{x}_k}$ can be written as $\mathbf{C}_{\mathbf{x}_k} = \sigma_k^2 \mathbf{a}_k \mathbf{a}_k^H + \mathbf{C}_{\mathbf{y}_k}$ where $\sigma_k^2 = \mathbf{w}_k^H \mathbf{C}_{\mathbf{x}_k}\mathbf{w}$, and the rank of $\mathbf{C}_{\mathbf{y}_k}$ is $d-1$, because $\mathbf{w}_k^H \mathbf{a}_k = 1$ due to the distortionless constraint, hence, $(\mathbf{I}_d - \mathbf{a}_k \mathbf{w}_k^H)\mathbf{a}_k = \mathbf{0}$.

For any vector $\mathbf{a}_k$, the couple of vectors $\mathbf{a}_{\alpha\mathrm{OC},k}(\mathbf{a}_k)$ and $\mathbf{w}_{\alpha,k}(\mathbf{a}_k)$ defined by (17) and (18), respectively, satisfy the OC, as shown in Section III-D. Therefore, the matrix $\frac{1}{\varsigma_k^2}\mathbf{C}_{\mathbf{x}_k} - \mathbf{a}_{\alpha\mathrm{OC},k}\mathbf{a}_{\alpha\mathrm{OC},k}^H$, where $\varsigma_k^2 = \mathbf{w}_{\alpha,k}^H \mathbf{C}_{\mathbf{x}_k}\mathbf{w}_{\alpha,k}$, has rank $d-1$. ■

ACKNOWLEDGMENT

The computational resources were provided by the e-INFRA CZ project (ID:90254).

REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications. Independent Component Analysis and Applications.* Amsterdam, Netherlands: Elsevier, 2010.

[2] K. Žmolíková, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Process. Mag.*, vol. 40, no. 3, pp. 8–29, May 2023.

[3] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[5] T. Kim, I. Lee, and T. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. 40th Asilomar Conf. Signals, Syst. Comput.*, Oct. 2006, pp. 1393–1396.

[6] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 550–563, Mar. 2010.

[7] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing.* Hoboken, NJ, USA: Wiley, 2002.

[8] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 185–189.

[9] D. Lahat and C. Jutten, "Joint independent subspace analysis using second-order statistics," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4891–4904, Sep. 2016.

[10] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-Gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, Feb. 2019.

[11] J. Eriksson and V. Koivunen, "Identifiability, separability, and uniqueness of linear ICA models," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 601–604, Jul. 2004.

[12] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, Apr. 2003, pp. 505–510.

[13] Z. Koldovský, V. Kautský, P. Tichavský, J. Čmejla, and J. Málek, "Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers," *IEEE Trans. Signal Process.*, vol. 69, pp. 2158–2173, 2021.

[14] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, Dec. 2001, pp. 722–727.

[15] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Proc. 14th Int. Workshop Image Anal. Multimedia Interactive Serv.*, 2013, pp. 1–4.

[16] Z. Wang, Y. Na, Z. Liu, Y. Li, B. Tian, and Q. Fu, "A semi-blind source separation approach for speech dereverberation," in *Proc. Interspeech* 2020, pp. 3925–3929.

[17] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 2417–2420.

[18] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki, "Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1032–1047, 2022.

[19] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[20] C. Hesse and C. James, "The FastICA algorithm with spatial constraints," *IEEE Signal Process. Lett.*, vol. 12, no. 11, pp. 792–795, Nov. 2005.

[21] S. Hirata et al., "Auxiliary-function-based steering vector estimation method for spatially regularized independent low-rank matrix analysis," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2024, pp. 1–6.

[22] R. Ikeshita and T. Nakatani, "Geometrically-regularized fast independent vector extraction by pure majorization-minimization," *IEEE Trans. Signal Process.*, vol. 72, pp. 1560–1575, 2024.

[23] S. Ma, X.-L. Li, N. M. Correa, T. Adali, and V. D. Calhoun, "Independent subspace analysis with prior information for fMRI data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 1922–1925.

[24] S. Bhinge, R. Mowakeaa, V. D. Calhoun, and T. Adalı, "Extraction of time-varying spatiotemporal networks using parameter-tuned constrained IVA," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1715–1725, Jul. 2019.

[25] A. Brendel, T. Haubner, and W. Kellermann, "A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, 2020.

[26] F. Nesta, S. Mosayyebpour, Z. Koldovský, and K. Paleček, "Audio/video supervised independent vector analysis through multimodal pilot dependent components," in *Proc. Eur. Signal Process. Conf.*, Sep. 2017, pp. 1190–1194.

[27] A. Hiroe, "Similarity-and-independence-aware beamformer with iterative casting and boost start for target source extraction using reference," *IEEE Open J. Signal Process.*, vol. 3, pp. 1–20, 2022.

[28] J. Gu, D. Yao, J. Li, and Y. Yan, "A novel semi-blind source separation framework towards maximum signal-to-interference ratio," *Signal Process.*, vol. 217, 2024, Art. no. 109359.

[29] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[30] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[31] K. Žmolìová et al., "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 800–814, Apr. 2019.

[32] R. Gu et al., "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. Interspeech*, 2019, pp. 4290–4294.

[33] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020.
[34] S. Zhang, J. Zhang, Y. Wang, and H. Yan, "Doa or speaker embedding: Which is better for multi-microphone target speaker extraction," *IEEE Signal Process. Lett.*, vol. 32, pp. 3350–3354, 2025.
[35] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[36] R. Carloni Gertosio, J. Bobin, and F. Acero, "Semi-blind source separation with learned constraints," *Signal Process.*, vol. 202, 2023, Art. no. 108776.
[37] N. Makishima et al., "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, Oct. 2019.
[38] F. Kang, F. Yang, and J. Yang, "Real-time independent vector analysis with a deep-learning-based source model," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 665–669.
[39] L. Li, H. Kameoka, and S. Makino, "FastMVAE2: On improving and accelerating the fast variational autoencoder-based source separation algorithm for determined mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 96–110, 2023.
[40] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6129–6133.
[41] N. Narisawa et al., "Independent deeply learned tensor analysis for determined audio source separation," in *Proc. 29th Eur. Signal Process. Conf.*, pp. 326–330, 2021.
[42] R. Scheibler and M. Togami, "Surrogate source model learning for determined source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 176–180.
[43] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 70–79, Jan. 2007.
[44] K. Matsumoto and K. Yatabe, "Determined BSS by combination of IVA and DNN via proximal average," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 871–875.
[45] R. Wang, L. Li, and T. Toda, "Dual-channel target speaker extraction based on conditional variational autoencoder and directional information," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1968–1979, 2024.
[46] R. Scheibler, W. Zhang, X. Chang, S. Watanabe, and Y. Qian, "End-to-end monaural multi-speaker ASR system without pretraining," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2023, pp. 496–501.
[47] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.
[48] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory.* Hoboken, NJ, USA: Wiley, 2002.
[49] Z. Koldovský, J. Málek, J. Čmejla, and S. O'Regan, "Informed FastICA: Semi-blind minimum variance distortionless beamformer," in *Proc. 18th Int. Workshop Acoustic Signal Enhancement*, 2024, pp. 95–99.
[50] Z. Koldovský, J. Málek, J. Čmejla, M. Vrátný, and W. Kellermann, "Fast algorithms for informed independent component/vector extraction," *EURASIP J. Adv. Signal Process.*, vol. 2025, Oct. 2025, Art. no. 56.
[51] Z. Koldovský, J. Čmejla, and S. O'Regan, "Blind capon beamformer based on independent component extraction: Single-parameter algorithm," *IEEE Signal Process. Lett.*, vol. 32, pp. 801–805, 2025.
[52] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, 1 st ed. Hoboken, NJ, USA: Wiley, 2018.
[53] T. Adalı, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal data fusion using source separation: Two effective models based on ICA and IVA and their properties," in *Proc. IEEE*, vol. 103, pp. 1478–1493, Sep. 2015.
[54] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, Sep. 2016.
[55] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, pp. 4–24, Apr. 1988.
[56] R. Ikeshita and T. Nakatani, "Independent vector extraction for joint blind source separation and dereverberation," vol. 28, pp. 972–976, 2021.
[57] V. Kautský, Z. Koldovský, P. Tichavský, and V. Zarzoso, "Cramér–Rao bounds for complex-valued independent component extraction: Determined and piecewise determined mixing models," *IEEE Trans. Signal Process.*, vol. 68, pp. 5230–5243, 2020.
[58] H. Li and T. Adalı, "Complex-valued adaptive signal processing using nonlinear functions," *EURASIP J. Adv. Signal Process.*, vol. 2008, Feb.2008, Art. no. 765615.
[59] K. B. Petersen and M. S. Pedersen, "The Matrix Cookbook," Oct. 2008. Version 20081110.
[60] M. Novey, T. Adali, and A. Roy, "A complex generalized Gaussian distribution—Characterization, generation, and estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1427–1433, Mar. 2010.
[61] J. Janský, Z. Koldovský, J. Málek, T. Kounovský, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP J. Audio, Speech, Music Process.*, vol. 2022, Jan. 2022, Art. no. 1.
[62] P. Tichavský, Z. Koldovský, and E. Oja, "Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 1189–1203, Apr. 2006.
[63] E. A. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Eindhoven, Netherlands, Tech. Rep, vol. 2, no. 2.4, p. 1, 2006.
[64] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proc. IEEE*, vol. 111, no. 5, pp. 465–499, May 2023.
[65] 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.
[66] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[67] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.