INFORMED FASTICA: SEMI-BLIND MINIMUM VARIANCE DISTORTIONLESS BEAMFORMER

Zbyněk Koldovský[1], *Jiří Málek*[1], *Jaroslav Čmejla*[1], *and Stephen O'Regan*[2]

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

ABSTRACT

Non-Gaussianity-based Independent Vector Extraction leads to the famous one-unit FastICA/FastIVA algorithm when the likelihood function is optimized using an approximate Newton-Raphson algorithm under the orthogonality constraint. In this paper, we replace the constraint with the analytic form of the minimum variance distortionless beamformer (MVDR), by which a semi-blind variant of FastICA/FastIVA is obtained. The side information here is provided by a weighted covariance matrix replacing the noise covariance matrix, the estimation of which is a frequent goal of neural beamformers. The algorithm thus provides an intuitive connection between model-based blind extraction and learning-based extraction. The algorithm is tested in simulations and speaker ID-guided speaker extraction, showing fast convergence and promising performance.

Index Terms— Blind Source Extraction, Independent Vector Analysis, Array Signal Processing, Noise Covariance Matrix, Minimum Variance Distortionless Beamformer

1. INTRODUCTION

We address the problem of extracting the source of interest (SOI) from a mixture of signals observed by multiple sensors. We particularly focus on speaker extraction from noisy multi-microphone recordings. Solutions to this problem can be achieved by beamforming or blind methods. On the one hand, beamforming offers optimal filters [1]. However, these can only be used if we know the key SOI or noise statistics, the precise estimation of which is as difficult a task as the problem itself. On the other hand, blind methods perform extraction based only on general assumptions such as the independence of SOI and other signals, as in Independent Vector Extraction (IVE) [2, 3, 4]. However, blind methods have limited accuracy and are burdened by uncertainties.

Both methodologies have, therefore, been adapted in various ways. Most recently, beamforming has been combined with deep neural networks that are trained to estimate the necessary SOI/noise covariance matrices [5, 6]. Blind methods are modified so that they can use side information to improve their performance, which is referred to as semi-blind, guided or informed methods. For example, in speaker extraction, the knowledge of the speaker's location is often used for constraining algorithms to enforce the extraction of the desired speaker [7, 8, 9]. There is also a significant class of methods that exploit information about the SOI activity, such as the variance profile of the speech or speaker's ID indicator [10, 11]. Although semi-blind methods form a diverse class of approaches, we observe in the literature that in many cases, a weighted covariance matrix of the input data plays a role, where the weights are computed using side information [12, 13, 14, 15]. Typically, the intention is that the weighted covariance matrix replaces the unknown noise/SOI covariance matrix: Semi-blind and beamforming methods meet here [16, 17].

In this paper, we propose an intuitive modification of the famous FastICA/FastIVE algorithm [18, 19] for semi-blind IVE. The original algorithm is derived using a fast second-order optimization of the likelihood function under the orthogonal constraint (OC). The OC enforces zero sample covariance between the estimated SOI and the other signals, which comes from the assumption of their independence. Analytically, the OC is equivalent to the minimum power distortionless (MPDR) beamformer, where the input-data covariance matrix and the mixing vector representing the relative acoustic transfer function (RTF) between the microphones and the SOI, play a role. The main idea here is to replace the covariance matrix in MPDR with its weighted counterpart, by which MPDR is replaced by an approximate MVDR (minimum variance distortionless beamformer), provided that the weighted covariance matrix approximates the noise covariance matrix. The weights are assumed to be provided through side information, which makes the algorithm semi-blind. Since MVDR is known to be less sensitive to the error in the mixing vector than MPDR, better global convergence to SOI can be expected. In simulations, we confirm this property, including improved accuracy, depending on how accurate the side information is. We also validate the algorithm on a speaker extraction task guided by deep training-based speaker identification.

The paper is organized as follows: The following section formulates the problem and recalls the MPDR and MVDR beamformers, as well as the basis for solving the problem using IVE. Section III describes the proposed algorithm and reveals its relations to the original FastICA. Section IV shows the results of the simulations and the speaker extraction experiment. Section V concludes the article.

2. PROBLEM FORMULATION

2.1. Observation and extraction model

We consider K mixtures of signals each observed through d sensors (microphones). The nth sample of the kth observed mixture is represented by a $d \times 1$ vector $\mathbf{x}^{[k]}(n)$ which contains the contribution of the source of interest (SOI) according to

$$\mathbf{x}^{[k]}(n) = \mathbf{a}^{[k]} s^{[k]}(n) + \mathbf{y}^{[k]}(n), \tag{1}$$

where $n = 1, \ldots, N$, $k = 1, \ldots, K$, $\mathbf{a}^{[k]}$ is the mixing vector of the SOI containing weights with which the SOI contributes to the sensors, and $\mathbf{y}^{[k]}(n)$ contains the other signals in the mixture. The

This work was supported by the Department of the Navy, Office of Naval Research Global, through Project No. N62909-23-1-2084.

extraction model assumes a separating vector $\mathbf{w}^{[k]}$ such that

$$s^{[k]}(n) \approx (\mathbf{w}^{[k]})^H \mathbf{x}^{[k]}(n) \tag{2}$$

in an exact or as close as possible sense. In the frequency-domain speaker extraction problem, k and n play the role of the frequency and frame index of the short-term Fourier transform, respectively. For simplicity, we omit writing the argument n further in the paper.

2.2. Optimum Beamformers

Array processing theory introduces optimum beamformers that seek $\mathbf{w}^{[k]}$ as a solution to an optimization problem [1]. In the context of this paper, we will consider two well-known beamformers. Minimum variance distortionless beamformer (MVDR) is the solution of

$$\mathbf{w}^{[k]}_{\mathrm{MVDR}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{C}^{[k]}_{\mathbf{y}} \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{a}^{[k]} = 1, \tag{3}$$

and minimum power distortionless beamformer (MPDR) is given by

$$\mathbf{w}^{[k]}_{\mathrm{MPDR}} = \arg\min_{\mathbf{w}} \mathbf{w}^H \mathbf{C}^{[k]}_{\mathbf{x}} \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{a}^{[k]} = 1. \tag{4}$$

$\mathbf{C}^{[k]}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}^{[k]}(\mathbf{y}^{[k]})^H]$ and $\mathbf{C}^{[k]}_{\mathbf{x}} = \mathrm{E}[\mathbf{x}^{[k]}(\mathbf{x}^{[k]})^H]$ denotes the covariance matrix of $\mathbf{y}^{[k]}$ and $\mathbf{x}^{[k]}$, respectively; $\mathrm{E}[\cdot]$ is the expectation operator. Hence, $\mathbf{w}^H \mathbf{C}^{[k]}_{\mathbf{y}} \mathbf{w}$ and $\mathbf{w}^H \mathbf{C}^{[k]}_{\mathbf{x}} \mathbf{w}$ is the power of the residual noise and of the overall output in $\mathbf{w}^H \mathbf{x}^{[k]}$, respectively. The distortionless constraint $\mathbf{w}^H \mathbf{a}^{[k]} = 1$ ensures that the SOI remains untouched in the beamformer's output. For nonsingular covariance matrices, it holds that

$$\mathbf{w}^{[k]}_{\mathrm{MVDR}} = \frac{(\mathbf{C}^{[k]}_{\mathbf{y}})^{-1} \mathbf{a}^{[k]}}{(\mathbf{a}^{[k]})^H (\mathbf{C}^{[k]}_{\mathbf{y}})^{-1} \mathbf{a}^{[k]}}, \tag{5}$$

$$\mathbf{w}^{[k]}_{\mathrm{MPDR}} = \frac{(\mathbf{C}^{[k]}_{\mathbf{x}})^{-1} \mathbf{a}^{[k]}}{(\mathbf{a}^{[k]})^H (\mathbf{C}^{[k]}_{\mathbf{x}})^{-1} \mathbf{a}^{[k]}}. \tag{6}$$

Practical deployments require knowledge of $\mathbf{a}^{[k]}$ and of the covariance matrices. MPDR is known to be sensitive to errors in the estimates of $\mathbf{C}^{[k]}_{\mathbf{x}}$ and $\mathbf{a}^{[k]}$ [1]. MVDR is less sensitive but it requires knowledge of $\mathbf{C}^{[k]}_{\mathbf{y}}$ or its estimate.

2.3. Independent Vector Extraction

IVE solves the extraction (and localization) problem blindly by jointly estimating $\mathbf{a}^{[k]}$ and $\mathbf{w}^{[k]}$ using only the observed signals. The mixture (1) is assumed to be determined[1], which means that $\mathbf{y}^{[k]}$ are assumed to belong to a $(d-1)$-dimensional subspace of so-called background sources represented by the $(d-1) \times 1$ vector $\mathbf{z}^{[k]}$. The *main* assumption is that $s^{[k]}$ is independent of $\mathbf{y}^{[k]}$ resp. $\mathbf{z}^{[k]}$, following the ICA model. The dependencies between the elements of the vector component $\mathbf{s} = [s^{[1]}, \ldots, s^{[K]}]^T$ are taken into account, following the IVA model [20].

The higher-order statistics-based source model assumes that the samples of signals are i.i.d. (independently and identically distributed); namely, $\mathbf{s}$ is distributed according to a multivariate non-Gaussian density, involving the dependencies among its elements. The pdf of $\mathbf{z}^{[k]}$ can be assumed circular Gaussian and uncorrelated across mixtures. This source model often neglects useful features of signals, nevertheless, it is simple and guarantees the identifiability of $\{\mathbf{w}^{[k]}, \mathbf{a}^{[k]}\}_{k=1}^K$ [21, 22].

[1]The determined case is convenient in terms of mathematical feasibility and is commonly used in practice where this condition is not satisfied.

Having N samples of observations for each k, the (quasi) maximum likelihood estimates $\{\hat{\mathbf{w}}^{[k]}, \hat{\mathbf{a}}^{[k]}\}_{k=1}^K$ can be obtained by finding the corresponding local maximum of

$$\begin{aligned}\mathcal{C}\left(\{\hat{\mathbf{w}}^{[k]}, \hat{\mathbf{a}}^{[k]}\}_{k=1}^K\right) = \hat{\mathrm{E}}\left[\log f\left(\left\{\frac{\hat{s}^{[k]}}{\hat{\sigma}_k}\right\}_{k=1}^K\right)\right] \\ - \sum_{k=1}^K \log \hat{\sigma}_k^2 - \sum_{k=1}^K \hat{\mathrm{E}}\left[(\hat{\mathbf{z}}^{[k]})^H (\mathbf{C}^{[k]}_{\mathbf{z}})^{-1} \hat{\mathbf{z}}^{[k]}\right] \\ + (d-2) \sum_{k=1}^K \log |\gamma^{[k]}|^2 + \text{const.},\end{aligned} \tag{7}$$

see, e.g., [23]. Here, $f(\cdot)$ is the model pdf of the SOI (a surrogate for the unknown true pdf) corresponding to a normalized random variable; $\hat{s}^{[k]} = (\hat{\mathbf{w}}^{[k]})^H \mathbf{x}^{[k]}$ is the extracted SOI given the current estimate of $\mathbf{w}^{[k]}$; $\hat{\sigma}_k^2 = (\hat{\mathbf{w}}^{[k]})^H \widehat{\mathbf{C}}^{[k]}_{\mathbf{x}} \hat{\mathbf{w}}^{[k]}$ is the sample-based variance of $\hat{s}^{[k]}$; $\hat{\mathbf{z}}^{[k]} = \hat{\mathbf{B}}^{[k]} \mathbf{x}^{[k]}$ are background signals estimates through the blocking matrix $\hat{\mathbf{B}}^{[k]} = [\hat{\mathbf{g}}^{[k]} \quad -\hat{\gamma}^{[k]} \mathbf{I}_{d-1}]$ where $\hat{\mathbf{a}}^{[k]} = \begin{bmatrix} \hat{\gamma}^{[k]} \\ \hat{\mathbf{g}}^{[k]} \end{bmatrix}$; $\hat{\mathrm{E}}[\cdot]$ denotes the sample-average operator; $\mathbf{C}^{[k]}_{\mathbf{z}}$ is the covariance matrix of $\mathbf{z}^{[k]}$, which is not known and must be replaced, e.g., by $\widehat{\mathbf{C}}^{[k]}_{\mathbf{z}} = \hat{\mathrm{E}}[\hat{\mathbf{z}}^{[k]}(\hat{\mathbf{z}}^{[k]})^H]$.

IVE involves indeterminacies, which are inherent (not only [6]) to blind methods: The role of the SOI can be played by any independent source in the mixture. It can also happen that a different source component (frequency) is extracted in each mixture, which is referred to as the permutation problem [24]. IVE alleviates this by taking into account the dependencies among the elements of $\mathbf{s}$. However, even if the permutation problem is solved, the global uncertainty of sources (speakers) remains.

This explains why it is necessary to look for the desired local maximum of (7). The optimization algorithm must be appropriately initialized and controlled. However, there is no way to ensure this on the basis of current assumptions. Additional information about the SOI is needed, which we assume in the following section.

3. PROPOSED ALGORITHM

3.1. Weighted covariance matrix

Let us assume that information about the SOI is available in the form of scalar signals $r_k(n)$, $k = 1, \ldots, K$. Typically, these signals can correspond to preliminary SOI estimates, activity indicators, estimated variance profiles and similarly so. The information can be joined for all k, in which case we have only one scalar signal $r(n) = r_1(n) = \cdots = r_K(n)$.

Then, let weighting functions $\alpha_k(\cdot)$, $k = 1, \ldots, K$, be given, which should be functions of $r_k(n)$. For simplicity, we will assume that the weighting functions are the same for all k, corresponding to the function denoted by $\alpha(\cdot)$. Then, we introduce weighted sample covariance matrices as

$$\widehat{\mathbf{C}}^{[k]}_{\alpha} = \hat{\mathrm{E}}\left[\alpha(r_k) \mathbf{x}^{[k]} \left(\mathbf{x}^{[k]}\right)^H\right], \quad k = 1, \ldots, K. \tag{8}$$

Most typically, $\alpha(\cdot)$ is a non-negative and non-increasing function such as, for example,

$$\alpha(r_k(n)) = \frac{1}{\epsilon + |r_k(n)|^2}, \tag{9}$$

where $\epsilon > 0$ is a small constant that prevents division by zero. If $r_k(n)$ somewhat indicates the activity of the SOI, the function (9)

applied in (8) is aimed to outweigh samples where the SOI is active (and vice versa) so that $\widehat{\mathbf{C}}_{\alpha}^{[k]}$ is as close as possible to $\mathbf{C}_{\mathbf{y}}^{[k]}$.

3.2. MVDR constraint

Couplings between the parametric vectors $\mathbf{a}^{[k]}$ and $\mathbf{w}^{[k]}$ can be applied to speed up or stabilize IVE algorithms. A popular one is the orthogonal constraint (OC), which enforces a zero sample-based correlation between $\hat{s}_k$ and $\hat{\mathbf{z}}^{[k]}$. Analytically, the OC is identical to MPDR given by (6), see [3]. So if $\hat{\mathbf{w}}^{[k]}$ is the dependent variable, the OC is given by

$$\hat{\mathbf{w}}_{\mathrm{OC}}^{[k]} = \frac{(\widehat{\mathbf{C}}_{\mathbf{x}}^{[k]})^{-1}\hat{\mathbf{a}}^{[k]}}{(\hat{\mathbf{a}}^{[k]})^{H}(\widehat{\mathbf{C}}_{\mathbf{x}}^{[k]})^{-1}\hat{\mathbf{a}}^{[k]}}, \tag{10}$$

where $\widehat{\mathbf{C}}_{\mathbf{x}}^{[k]} = \hat{\mathrm{E}}[\mathbf{x}^{[k]}(\mathbf{x}^{[k]})^{H}]$. Therefore, orthogonally constrained IVE algorithms are seeking the optimum of $\mathcal{C}(\{\hat{\mathbf{w}}_{\mathrm{OC}}^{[k]}, \hat{\mathbf{a}}^{[k]}\}_{k=1}^{K})$ with respect to $\{\hat{\mathbf{a}}^{[k]}\}_{k=1}^{K}$ and can thus be seen as blind MPDR beamformers [25].

Motivated by this interpretation, we propose to replace (10) by

$$\hat{\mathbf{w}}_{\alpha}^{[k]} = \frac{(\widehat{\mathbf{C}}_{\alpha}^{[k]})^{-1}\hat{\mathbf{a}}^{[k]}}{(\hat{\mathbf{a}}^{[k]})^{H}(\widehat{\mathbf{C}}_{\alpha}^{[k]})^{-1}\hat{\mathbf{a}}^{[k]}}. \tag{11}$$

Provided that $\widehat{\mathbf{C}}_{\alpha}^{[k]}$ is an estimate of $\mathbf{C}_{\mathbf{y}}^{[k]}$, then (11) could be seen as an estimate of the MVDR beamformer given by (5).

3.3. Second-order algorithm

We now propose an algorithm that seeks the optimum point of $\mathcal{C}(\{\hat{\mathbf{w}}_{\alpha}^{[k]}, \hat{\mathbf{a}}^{[k]}\}_{k=1}^{K})$ with respect to $\{\hat{\mathbf{a}}^{[k]}\}_{k=1}^{K}$, which can be interpreted as a semi-blind MVDR beamformer since $\widehat{\mathbf{C}}_{\alpha}^{[k]}$ provide side information. The optimization approach is heuristically derived based on the approximate Newton-Raphson update [23]

$$\hat{\mathbf{a}}^{[k]} \leftarrow \hat{\mathbf{a}}^{[k]} - (\mathbf{H}^{[k]*})^{-1}\boldsymbol{\Delta}^{[k]}, \quad k = 1, \ldots, K, \tag{12}$$

where $(\cdot)^{*}$ denotes the conjugate value, and

$$\boldsymbol{\Delta}^{[k]} = \frac{\partial}{\partial(\mathbf{a}^{[k]})^{*}}\mathcal{C}(\{\hat{\mathbf{w}}_{\alpha}^{[k]}, \hat{\mathbf{a}}^{[k]}\}_{k=1}^{K}), \tag{13}$$

$$\mathbf{H}^{[k]} = \left.\frac{\partial(\boldsymbol{\Delta}^{[k]})^{T}}{\partial\mathbf{a}^{[k]}}\right|_{\hat{\mathbf{a}}^{[k]}=\mathbf{a}^{[k]}, N\to+\infty}. \tag{14}$$

After some manipulations, we arrive at the following update rule

$$\hat{\mathbf{a}}^{[k]} \leftarrow \hat{\mathbf{a}}^{[k]} - \frac{\hat{\nu}_k}{\hat{\nu}_k - \hat{\rho}_k}\frac{\hat{\sigma}_{\alpha,k}^{2}}{\hat{\sigma}_k^{2}}\left(\hat{\mathbf{a}}^{[k]} - \hat{\nu}_k^{-1}\hat{\mathrm{E}}\left[\phi_k\left(\overline{\mathbf{s}}\right)\frac{\mathbf{x}^{[k]}}{\hat{\sigma}_k}\right]\right), \tag{15}$$

where $\phi_k = -\frac{\partial}{\partial s_k}\log f$, $\hat{\nu}_k = \hat{\mathrm{E}}\left[\frac{\hat{s}_k}{\hat{\sigma}_k}\phi_k\left(\overline{\mathbf{s}}\right)\right]$, $\hat{\rho}_k = \hat{\mathrm{E}}\left[\frac{\partial\phi_k}{\partial s_k^{*}}\left(\overline{\mathbf{s}}\right)\right]$, $\hat{\sigma}_{\alpha,k}^{2} = (\hat{\mathbf{w}}_{\alpha}^{[k]})^{H}\widehat{\mathbf{C}}_{\alpha}^{[k]}\hat{\mathbf{w}}_{\alpha}^{[k]}$, and $\overline{\mathbf{s}} = [\frac{\hat{s}_1}{\hat{\sigma}_1}, \ldots, \frac{\hat{s}_K}{\hat{\sigma}_K}]^{T}$ is the normalized SOI vector component. Detailed derivations are not given in this paper due to limited space; see the Appendix.

3.4. Relation to the FastICA algorithm

Let us consider the case $K = 1$, in which case the index k can be omitted from the notation. We can easily represent a situation where $r(n)$ does not bring any useful information, e.g., by putting it equal to a constant such that $\alpha = 1$. Then, $\widehat{\mathbf{C}}_{\alpha} = \widehat{\mathbf{C}}_{\mathbf{x}}$ and $\hat{\sigma}_{\alpha}^{2} = \hat{\sigma}^{2}$. We can assume that the output signal is normalized so that $\hat{\sigma}^{2} = 1$. Then, (15) simplifies to

$$\hat{\mathbf{a}} \leftarrow \hat{\mathbf{a}} - \frac{\hat{\nu}}{\hat{\nu} - \hat{\rho}}\left(\hat{\mathbf{a}} - \hat{\nu}^{-1}\hat{\mathrm{E}}[\phi(\hat{s})\mathbf{x}]\right). \tag{16}$$

According to the constraint (11), which now coincides with (10), we can apply substitution $\hat{\mathbf{a}} = \widehat{\mathbf{C}}_{\mathbf{x}}\hat{\mathbf{w}}$ and rewrite (16) to the form

$$\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} - \frac{\hat{\nu}}{\hat{\nu} - \hat{\rho}}\left(\hat{\mathbf{w}} - \hat{\nu}^{-1}\widehat{\mathbf{C}}_{\mathbf{x}}^{-1}\hat{\mathrm{E}}[\phi(\hat{s})\mathbf{x}]\right). \tag{17}$$

Since the scale of the output signal (thus also of the parameter vector $\mathbf{w}$) can be arbitrary, the right-hand side of (17) can be multiplied by the scalar $\hat{\rho} - \hat{\nu}$, which results in the well known FastICA update rule [18]

$$\hat{\mathbf{w}} \leftarrow \hat{\rho}\hat{\mathbf{w}} - \widehat{\mathbf{C}}_{\mathbf{x}}^{-1}\hat{\mathrm{E}}[\phi(\hat{s})\mathbf{x}]. \tag{18}$$

The update (15) thus can be seen as an informed extension of the FastICA algorithm.

4. EXPERIMENTAL VALIDATION

4.1. Simulations

In simulations, we generate artificial mixtures and extract the SOI to verify the performance of the proposed algorithm and compare it with the original one-unit FastICA[2]. The algorithms are compared in two regimes: (1) mixtures are processed separately as if $K = 1$ (the original and proposed method are denoted by "FastICA" and "iFastICA", respectively) and jointly as if $K = 6$ (the methods are denoted "FastIVA" and "iFastIVA", respectively). We primarily focus on their global convergence, i.e., whether the algorithms extracted the desired source. Each extracted signal is evaluated by means of signal-to-interference ratio (SIR); the success rate of an algorithm is defined as the percentage of trials where it achieves SIR higher than 3dB. The SIR is then averaged over these "successful" trials, which reflects the accuracy of the algorithm.

In a trial of the simulation, a complex-valued mixture of $d = 5$ super-Gaussian signals of length N is generated. The first column of the random mixing matrix is the mixing vector of the SOI, which is the first generated signal. The initial SIR ($\mathrm{SIR}_{\mathrm{ini}}$), defined as the ratio of its scale to the mean scale of the other signals, is set to the chosen value through a corresponding re-scaling of the SOI. The side information in iFastICA/iFastIVA is provided through $r_k(n) = \sqrt{1-\epsilon^2}s^{[k]}(n) + \epsilon v^{[k]}(n)$, where $v^{[k]}(n)$ is standard Gaussian, and $\epsilon^2 = 0.5$; the weighting function is given by (9). The algorithms are initialized randomly in a random vicinity of the SOI.

Fig. 1 shows the success rate and average SIR as functions of N (when $\mathrm{SIR}_{\mathrm{ini}} = 0$ dB) and $\mathrm{SIR}_{\mathrm{ini}}$ (when $N = 200$). The success rate shows that the proposed semi-blind methods significantly profit from the side information compared to the blind algorithms. The accuracy of the semi-blind methods is, on the other hand, similar to that of the blind ones or even slightly worse. The semi-blind methods thus bring improvements in terms of global convergence.

Of particular interest are the challenging values of $N = 10, \ldots, 100$ and $\mathrm{SIR}_{\mathrm{ini}} = -20, \ldots, -5$ dB. The success rate of blind algorithms here is often less than 10%, pointing to the problem of blind extraction if the data is too short or if the SOI is too weak. The semi-blind methods offer a solution in this regard. Finally, the performance of iFastIVA/FastIVA is higher than that of

[2]We use the implementation from [26] with $K = T = L = 1$.

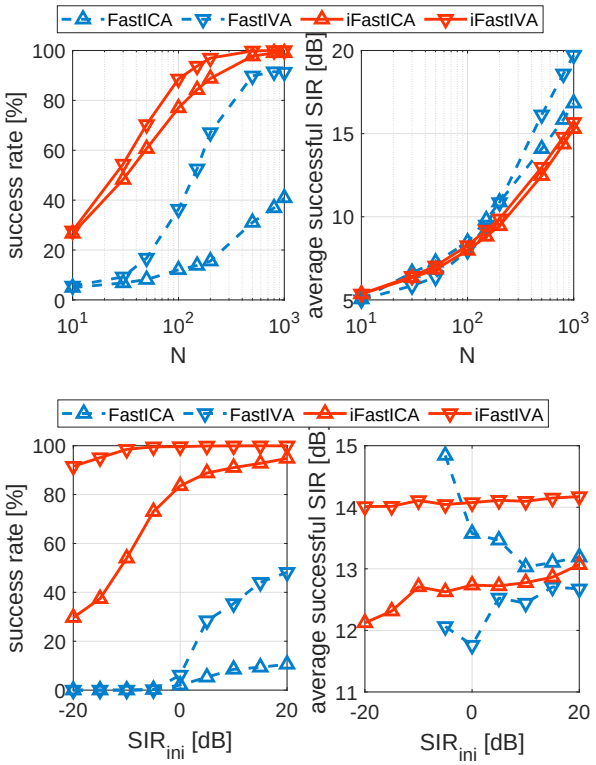

Fig. 1. Success rate and SIR averaged over successful trials of the compared algorithms as functions of N (when $\mathrm{SIR}_{\mathrm{ini}} = 0$ dB) and $\mathrm{SIR}_{\mathrm{ini}}$ (when $N = 200$); each setting was repeated in 1000 trials.

iFastICA/FastICA, which clearly demonstrates the benefit of joint extraction, also in the case of the semi-blind methods.

Simulations can be used to test the algorithms depending on other key parameters such as ϵ^2. The code is publicly available[3].

4.2. Target speaker extraction

FastIVA and iFastIVA are compared to the informed auxiliary function-based iAuxIVE (alternative naming for the piloted AuxIVE from [15]) in the frequency-domain speaker extraction problem. Reverberated mixtures of two speakers contained in the multi-channel Wall Street Journal dataset (MC-WSJ0-2mix, [27]) are considered. The dataset contains $3,000$ simulated mixtures; we use the first four microphones. Each mixture contains two active speakers; thus, there are $6,000$ extraction experiments in total. The sources are mixed at SIR between $\langle -5, +5 \rangle$ dB. The recordings are highly reverberant ($T_{60} \in \langle 200, 600 \rangle$ ms). The sampling frequency is 16 kHz; the STFT window length is 1000 samples; the window shift is 200.

The prior information is obtained using speaker identification via embeddings. The embeddings are computed in the same manner as described for the MC-WSJ0-2mix in [11]. The signal $r(n)$ (independent of k) and the pilot signal for the iAuxIVE are equal to the energy of the mixture when the target speaker appears to be dominant and zero otherwise. The target is assumed dominant within a closed set of speakers when its reference embedding is the most similar to the embedding computed using the nth frame of the mixture.

Fig. 2 presents the results evaluated using Signal-To-Interference

[3]https://github.com/koldovsk/IWAENC-2024-Informed-FastICA

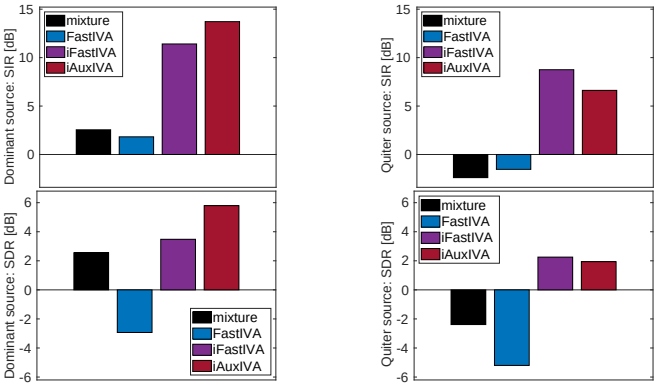

Fig. 2. MC-WSJ0-2mix: Speaker extraction metrics achieved by the compared algorithms.

Ratio (SIR) and Signal-To-Distortion Ratio (SDR) defined by the BSS_EVAL [28] toolbox. The experiment is evaluated separately for the dominant and the quieter speaker. FastIVA fails here as it cannot distinguish the target, whereas both informed algorithms are able to extract it. The iAuxIVE algorithm is more successful when extracting the dominant source, whereas iFastIVA excels at extracting the quieter speaker.

5. CONCLUSIONS

We have proposed semi-blind algorithms for source extraction that can exploit side information to reach the desired source. The algorithms combine IVE and MVDR intuitively. Future work should focus on deeper theoretical analysis to provide insight into which type of side information can be used most effectively.

6. APPENDIX

The computation of (13), after ϕ_k is divided by $\hat{\nu}_k$ (see [23] for justification of this step), gives

$$\boldsymbol{\Delta}^{[k]} = \sigma^2_{\alpha,k}(\widehat{\mathbf{C}}^{[k]}_{\alpha})^{-1}\left(\hat{\mathbf{a}}^{[k]} - \hat{\nu}_k^{-1}\hat{\mathrm{E}}\left[\phi_k(\bar{\mathbf{s}}) \cdot \frac{\mathbf{x}^{[k]}}{\sigma_k}\right]\right) + \begin{pmatrix} (\hat{\mathbf{g}}^{[k]})^H(\widehat{\mathbf{C}}^{[k]}_{\mathbf{z}})^{-1}\mathbf{q}^{[k]} \\ -(\hat{\gamma}^{[k]})^*(\widehat{\mathbf{C}}^{[k]}_{\mathbf{z}})^{-1}\mathbf{q}^{[k]} \end{pmatrix}, \quad (19)$$

where $\mathbf{q}^{[k]} = \hat{\mathrm{E}}[\hat{\mathbf{z}}^{[k]}\hat{s}^{[k]}]$, which is the sample covariance of the current SOI estimate and the background signals. For the orthogonal constraint (10), $\mathbf{q}^{[k]}$ equals zero. To ensure this, $\hat{\mathbf{a}}^{[k]}$ should be recomputed after (11) as $\hat{\mathbf{a}}^{[k]} \leftarrow \hat{\sigma}_k^{-2}\widehat{\mathbf{C}}^{[k]}_{\mathbf{x}}\hat{\mathbf{w}}^{[k]}$, which is equivalent to (6). Next, in the computation of (14), we neglect all rank-one terms, which results in

$$\mathbf{H}^{[k]} = \hat{\sigma}^2_{\alpha,k}\left(\mathbf{I}_d - \frac{\hat{\rho}_k}{\hat{\nu}_k}\frac{\hat{\sigma}^2_{\alpha,k}}{\hat{\sigma}^2_k}(\widehat{\mathbf{C}}^{[k]*}_{\alpha})^{-1}\widehat{\mathbf{C}}^{[k]*}_{\mathbf{x}}\right)(\widehat{\mathbf{C}}^{[k]*}_{\alpha})^{-1}. \quad (20)$$

Then, we apply the approximation $\frac{\hat{\sigma}^2_{\alpha,k}}{\hat{\sigma}^2_k}(\widehat{\mathbf{C}}^{[k]}_{\alpha})^{-1}\widehat{\mathbf{C}}^{[k]}_{\mathbf{x}} \approx \mathbf{I}_d$ and multiply $\mathbf{H}^{[k]}$ by an experimentally verified factor $\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{\alpha,k}}$. By putting into (12), (15) follows. ■

7. REFERENCES

[1] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, John Wiley & Sons, Inc., 2002.

[2] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, Wiley Publishing, 1st edition, 2018.

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

[4] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 185–189.

[5] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 196–200.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[8] A. H. Khan, M. Taseska, and E. A. P. Habets, *A Geometrically Constrained Independent Vector Analysis Algorithm for Online Source Extraction*, pp. 396–403, Springer International Publishing, Cham, 2015.

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

[10] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2417–2420.

[11] J. Malek, J. Jansky, Z. Koldovsky, T. Kounovsky, J. Cmejla, and J. Zdansky, "Target speech extraction: Independent vector extraction guided by supervised speaker identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2295–2309, 2022.

[12] B. J. Cho, J.-M. Lee, and H.-M. Park, "A beamforming algorithm based on maximum likelihood of a complex gaussian distribution with time-varying variances for robust speech recognition," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1398–1402, 2019.

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

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

[15] J. Janský, Z. Koldovský, J. Málek, T. Kounovský, and J. Čmejla, "Auxiliary function-based algorithm for blind extraction of a moving speaker," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1, Jan 2022.

[16] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3461–3475, 2021.

[17] 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. on Audio, Speech, and Language Processing*, vol. 30, pp. 1032–1047, 2022.

[18] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[19] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[20] T. Kim, I. Lee, and T. Lee, "Independent vector analysis: Definition and algorithms," in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, Oct 2006, pp. 1393–1396.

[21] M. Anderson, G. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4399–4410, Sept 2014.

[22] V. Kautský, Z. Koldovský, and P. Tichavský, "Performance bound for blind extraction of non-Gaussian complex-valued vector component from Gaussian background," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, May 2019, vol. 5287–5291.

[23] 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 Transactions on Signal Processing*, vol. 69, pp. 2158–2173, 2021.

[24] 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," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[25] Z. Koldovský, P. Tichavský, and V. Kautský, "Orthogonally constrained independent component extraction: Blind MPDR beamforming," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1155–1159.

[26] Z. Koldovský, V. Kautský, and P. Tichavský, "Double non-stationarity: Blind extraction of independent nonstationary vector/component from nonstationary mixtures—Algorithms," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5102–5116, 2022.

[27] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP 2018*. IEEE, 2018, pp. 1–5.

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.