EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Auxiliary function-based algorithm for blind extraction of a moving speaker

Jakub Janský*, Zbyněk Koldovský, Jiří Málek, Tomáš Kounovský and Jaroslav Čmejla

Abstract

In this paper, we propose a novel algorithm for blind source extraction (BSE) of a moving acoustic source recorded by multiple microphones. The algorithm is based on independent vector extraction (IVE) where the contrast function is optimized using the auxiliary function-based technique and where the recently proposed constant separating vector (CSV) mixing model is assumed. CSV allows for movements of the extracted source within the analyzed batch of recordings. We provide a practical explanation of how the CSV model works when extracting a moving acoustic source. Then, the proposed algorithm is experimentally verified on the task of blind extraction of a moving speaker. The algorithm is compared with state-of-the-art blind methods and with an adaptive BSE algorithm which processes data in a sequential manner. The results confirm that the proposed algorithm can extract the moving speaker better than the BSE methods based on the conventional mixing model and that it achieves improved extraction accuracy than the adaptive method.

Keywords: Blind source separation, Independent vector analysis, Speaker extraction, Speech enhancement, Moving sources

1 Introduction

This paper addresses the problem when sound is sensed by multiple microphones and the goal is to extract a signal of interest originating from an individual source. We particularly address the case when the corresponding source is a speaker which is moving during the recording. Unknown situation is considered where no information about the environment and the positions of microphones and sources is available and no training data are available. This is the task of blind source separation (BSS), or particularly, of blind source extraction (BSE). These signal processing fields embrace numerous methods such as nonnegative matrix/tensor factorization, clustering and classification approaches, or sparsity-awareness methods; see [1–3] for surveys. We will consider the approach of independent component analysis (ICA) where signals are separated into original signals based on the assumption that the original signals are *statistically independent* [4]. In case of audio sources such as speakers, this fundamental condition is met, which makes ICA attractive for practical applications.

ICA can separate instantaneous mixtures of non-Gaussian independent signals up to their indeterminable order and scales [5]. Since acoustic mixtures are convolutive due to delays and reverberation, the narrow-band approach can be considered. Here, ICA is applied in the short-time Fourier transform (STFT) domain separately in each frequency bin; the approach referred to as frequency-domain ICA (FDICA) [3, 6]. However, the separate applications of ICA in FDICA cause the so-called *permutation problem* due to the indeterminable order of separated signals: The separated frequency components have a random order and must be aligned in order to retrieve the full-band separated signals [7]. Independent vector analysis (IVA) treats all frequencies simultaneously using a joint statistical source model [8, 9]. The frequency components of the original signals form the so-called vector components. IVA aims at maximizing higher-order dependencies between the frequency components within each vector component while the whole vector compo-

*Correspondence: jakub.jansky@tul.cz
Acoustic Signal Analysis and Processing Group, Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, Liberec, Czech Republic

nents should be independent [9]. IVA is thus an extension of ICA to joint separation of several instantaneous mixtures (one per frequency bin).

A recent extension of IVA is independent low-rank matrix analysis (ILRMA) where the vector components are assumed to obey a low-rank source model. For example, ILRMA combines IVA and nonnegative matrix eactorization (NMF) in [10, 11] and involves deep learning in [12]. The counterparts of ICA and IVA designed for BSE, i.e., for the extraction of one independent source, are called independent component/vector extraction (ICE/IVE) [13, 14]. Very recently, IVE has been extended towards simultaneous source extraction and dereverberation [15].

In principle, the aforementioned methods differ in source modeling while they share the conventional time-invariant linear mixing model. This model describes situations that are not changing during the recording time, which also means that the sources-speakers are assumed to be static. To separate/extract moving sources, the methods can be used in an adaptive way by being applied on short intervals during which the mixture is approximately static. Such modifications are typically implemented to process data sample-by-sample (frame-by-frame) or batch-by-batch using some forgetting update of inner parameters [16–18]; many such methods have been considered also in biomedical applications; see, e.g., [19]. Although these methods are useful, they have several shortcomings. Namely, the sources can be separated in a different order at different times due to the indeterminacy of ICA; we refer to this as to the *discontinuity problem*. Also, the separation accuracy is limited by the length of context from which the time-variant separating parameters are computed. The methods involve parameters such as learning rate or forgetting factors for recursive processing. Optimum values of those parameters depend on input data in an unknown way. The control and tuning of these adaptive implementations, therefore, poses a difficult and application-dependent problem.

In this paper, we propose a novel algorithm for IVE based on the constant separating vector (CSV) mixing model, which is called CSV-AuxIVE. CSV-AuxIVE belongs to the family of auxiliary function-based methods [17, 20, 21]. These methods use a majorization-minimization approach for finding the optimum of a contrast function derived based on the maximum likelihood principle and do not involve any learning rate parameter. In particular, CSV-AuxIVE could be seen as an extension of the recent OverIVA algorithm from [22] allowing for the CSV mixing model. CSV has been first considered in the preliminary conference report [23]. It involves time-variant mixing parameters while it simultaneously assumes time-invariant (constant) separating parameters. The model enables us to avoid the discontinuity problem and to improve the extraction performance because the extraction accuracy depends on the length of the entire recording modeled by CSV [24]. The proposed CSV-AuxIVE adopts these important features and provides a new blind method, which is much faster than the gradient-based algorithm used in [23].

The article is organized as follows: in Section 2, the technical definition of the BSE problem is given, the CSV mixing model is described and explained from a practical point of view, and the contrast function for the blind extraction is derived. In Section 3, the proposed CSV-AuxIVE algorithm is described, including its piloted variant that enables a partial control of convergence using prior knowledge of the desired signal. Section 4 is devoted to experimental evaluations based on simulated as well as real-world data. The paper is concluded in Section 5. A supplementary material to this paper contains a detailed derivation of the gradient-based algorithm from [23] referred to as BOGIVE$_{\mathbf{w}}$.

Notation Plain letters denote scalars, bold letters denote vectors, and bold capital letters denote matrices. Upper indices such as $\cdot^T$, $\cdot^H$, or $\cdot^*$ denote, respectively, transposition, conjugate transpose, or complex conjugate. The Matlab convention for matrix/vector concatenation and indexing will be used, e.g., $[\,1;\,\mathbf{g}] = \left[1,\,\mathbf{g}^T\right]^T$ and $(\mathbf{a})_i$ is the ith element of $\mathbf{a}$. $\mathrm{E}[\,\cdot]$ stands for the expectation operator, and $\hat{\mathrm{E}}[\,\cdot]$ is the average taken over all available samples of the symbolic argument. The letters k and t are used as integer indices of frequency bin and block, respectively; $\{\cdot\}_k$ is a short notation of the argument with all values of index k, e.g., $\{\mathbf{w}_k\}_k$ means $\mathbf{w}_1, \ldots, \mathbf{w}_K$, and $\{\mathbf{w}_{k,t}\}_{k,t}$ means $\mathbf{w}_{1,1}, \ldots, \mathbf{w}_{K,T}$.

2 Problem formulation

A static mixture of audio signals that propagate in an acoustic environment from point sources to microphones can be described by the time-invariant convolutive model. Let there be d sources observed by m microphones. The signal on the ith microphone is described by

$$x_i(n) = \sum_{j=1}^{d}\sum_{\tau=0}^{L-1} h_{ij}(\tau)s_j(n-\tau), \quad i = 1, \ldots, m, \tag{1}$$

where n is the sample index, $s_1(n), \ldots, s_d(n)$ are the original signals coming from the sources, and h_{ij} denotes the time-invariant impulse response between the jth source and ith microphone of length L.

In the short-time Fourier transform (STFT) domain, convolution can be approximated by multiplication. Let $x_i(k,\ell)$ and $s_j(k,\ell)$ denote, respectively, the STFT coefficient of $x_i(n)$ and $s_j(n)$ at frequency k and frame ℓ. Then, (1) can be replaced by a set of K complex-valued linear instantaneous mixtures

$$\mathbf{x}_k = \mathbf{A}_k \mathbf{s}_k, \qquad k = 1, \ldots, K, \tag{2}$$

where $\mathbf{x}_k$ and $\mathbf{s}_k$ are symbolic vectors representing, respectively, $[x_1(k,\ell), \ldots, x_m(k,\ell)]^T$ and $[s_1(k,\ell), \ldots, s_d(k,\ell)]^T$, for any frame $\ell = 1, \ldots, N$; $\mathbf{A}_k$ stands for the $m \times d$ mixing matrix whose ijth element is related to the kth Fourier coefficient of the impulse response h_{ij}; K is the frequency resolution of the STFT; for detailed explanations, see, e.g., Chapters 1 through 3 in [3].

2.1 Blind source extraction

For the BSE problem, we can write (2) in the form

$$\mathbf{x}_k = \mathbf{a}_k s_k + \mathbf{y}_k, \qquad k = 1, \ldots, K, \tag{3}$$

where s_k represents the *source of interest* (SOI), $\mathbf{a}_k$ is the corresponding column of $\mathbf{A}_k$, called the *mixing vector*, and $\mathbf{y}_k$ represents the remaining signals in $\mathbf{x}_k$, i.e., $\mathbf{y}_k = \mathbf{x}_k - \mathbf{a}_k s_k$.

Since there is the ambiguity that any of the original sources can play the role of the SOI, we can assume, without loss of generality, that the SOI corresponds to the first source in (2); hence, $\mathbf{a}_k$ is the first column of $\mathbf{A}_k$. The problem of guaranteeing the extraction of the desired SOI will be addressed in Section 3.3.

The assumption that the original signals in (2) are independent implies that s_k and $\mathbf{y}_k$ are independent. We will also assume that $m = d$, i.e., that there is the same number of microphones as that of the sources. It follows that the mixing matrices $\mathbf{A}_k$ are square. By assuming also that they are non-singular[1] and that their inverse matrices exist, the existence of a *separating vector* $\mathbf{w}_k$ (the first row of $\mathbf{A}_k^{-1}$) such that $\mathbf{w}_k^H \mathbf{x}_k = s_k$ is guaranteed. We pay for this advantage by the limitation that $\mathbf{y}_k$ belongs to a subspace of dimension $d-1$. In other words, the covariance of $\mathbf{y}_k$ is assumed to have rank $d-1$ as opposed to real recordings where the typical rank is d (e.g. due to sensor and environment noises). Nevertheless, the assumption $m = d$ brings more advantages than disadvantages as shown in [10]. One way to compensate is to increase the number of microphones so that the ratio $\frac{d-1}{d}$ approaches 1. BSE appears to be computationally more efficient than BSS when d is large since, in BSE, $\mathbf{y}_k$ is not separated into individual signals.

In [13], the BSE problem is formulated by exploiting the fact that the $d-1$ latent variables (background signals) involved in $\mathbf{y}_k$ can be defined arbitrarily. An effective parameterization that involves only the mixing and separating vectors related to the SOI has been derived. Specifically, $\mathbf{A}_k$ and $\mathbf{A}_k^{-1}$ (denoted as $\mathbf{W}_k$) have the structure

$$\mathbf{A}_k = \begin{pmatrix} \mathbf{a}_k & \mathbf{Q}_k \end{pmatrix} = \begin{pmatrix} \gamma_k & \mathbf{h}_k^H \\ \mathbf{g}_k & \frac{1}{\gamma_k}(\mathbf{g}_k \mathbf{h}_k^H - \mathbf{I}_{d-1}) \end{pmatrix}, \tag{4}$$

and

$$\mathbf{W}_k = \begin{pmatrix} \mathbf{w}_k^H \\ \mathbf{B}_k \end{pmatrix} = \begin{pmatrix} \beta_k{}^* & \mathbf{h}_k^H \\ \mathbf{g}_k & -\gamma_k \mathbf{I}_{d-1} \end{pmatrix}, \tag{5}$$

where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix, $\mathbf{w}_k$ denotes the separating vector which is partitioned as $\mathbf{w}_k = [\beta_k; \mathbf{h}_k]$; the mixing vector $\mathbf{a}_k$ is partitioned as $\mathbf{a}_k = [\gamma_k; \mathbf{g}_k]$. The vectors $\mathbf{a}_k$ and $\mathbf{w}_k$ are linked through the so-called *distortionless constraint* $\mathbf{w}_k^H \mathbf{a}_k = 1$, which, equivalently, means

$$\beta_k^* \gamma_k + \mathbf{h}_k^H \mathbf{g}_k = 1, \qquad k = 1, \ldots, K. \tag{6}$$

$\mathbf{B}_k = [\mathbf{g}_k, -\gamma_k \mathbf{I}_{d-1}]$ is called the *blocking matrix* as it satisfies that $\mathbf{B}_k \mathbf{a}_k = \mathbf{0}$. The background signals are given by $\mathbf{z}_k = \mathbf{B}_k \mathbf{x}_k = \mathbf{B}_k \mathbf{y}_k$, and it holds that $\mathbf{y}_k = \mathbf{Q}_k \mathbf{z}_k$. To summarize, (2) is recasted for the BSE problem as

$$\mathbf{x}_k = \begin{pmatrix} \gamma_k & \mathbf{h}_k^H \\ \mathbf{g}_k & \frac{1}{\gamma_k}(\mathbf{g}_k \mathbf{h}_k^H - \mathbf{I}_{d-1}) \end{pmatrix} \begin{pmatrix} s_k \\ \mathbf{z}_k \end{pmatrix}, \quad k = 1, \ldots, K. \tag{7}$$

2.2 CSV mixing model

Now, we turn to an extension of (7) to time-varying mixtures. Let the available samples of the observed signals (meaning the STFT coefficients from N frames) be divided into T intervals; for the sake of simplicity, we assume that the intervals have the same integer length $N_b = N/T$. The intervals will be called blocks and will be indexed by $t \in \{1, \ldots, T\}$.

A straightforward extension of (7) to time-varying mixtures is when all parameters, i.e., the mixing and separating vectors, are block-dependent. However, such an extension brings no advantage compared to processing each block separately. In the constant separating vector (CSV) mixing model, it is assumed that only the mixing vectors are block-dependent while the separating vectors are constant over the blocks. Hence, the mixing and de-mixing matrices on the tth block are parameterized, respectively, as

$$\mathbf{A}_{k,t} = \begin{pmatrix} \mathbf{a}_{k,t} & \mathbf{Q}_{k,t} \end{pmatrix} = \begin{pmatrix} \gamma_{k,t} & \mathbf{h}_k^H \\ \mathbf{g}_{k,t} & \frac{1}{\gamma_{k,t}}(\mathbf{g}_{k,t} \mathbf{h}_k^H - \mathbf{I}_{d-1}) \end{pmatrix}, \tag{8}$$

and

$$\mathbf{W}_{k,t} = \begin{pmatrix} \mathbf{w}_k^H \\ \mathbf{B}_{k,t} \end{pmatrix} = \begin{pmatrix} \beta_k^* & \mathbf{h}_k^H \\ \mathbf{g}_{k,t} & -\gamma_{k,t} \mathbf{I}_{d-1} \end{pmatrix}. \tag{9}$$

Each sample of the observed signals on the tth block is modeled according to

$$\mathbf{x}_{k,t} = \mathbf{A}_{k,t} \begin{pmatrix} s_{k,t} \\ \mathbf{z}_{k,t} \end{pmatrix}, \tag{10}$$

[1] This assumption simplifies the theoretical development of algorithms and does not hamper the applicability of the methods on real signals. For example, practical recordings always contain some noise and so behave as mixtures with a non-singular mixing matrix.

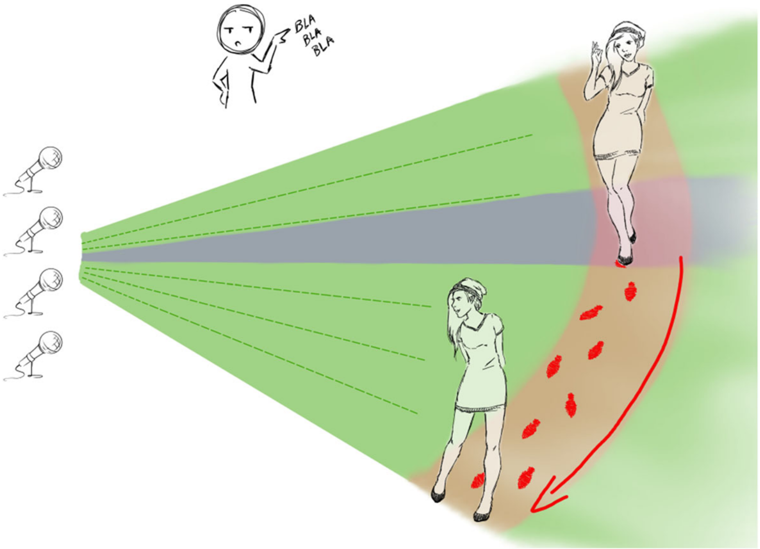


Fig. 1 How the blind extraction of a moving speaker can be solved based on CSV. The narrow area (in gray) stands for a typical focus of a separating filter obtained by the static mixing models. It is able to extract the speaker only from a particular position. The green area denotes the focus of a separating filter obtained through CSV: it covers the entire area of the speaker's movement during the recording. Such separating vector exists because there is a sufficient number of microphones and because the interfering speaker is located outside the area of the target speaker

where $s_{k,t}$ and $\mathbf{z}_{k,t}$ represent, respectively, the kth frequency of the SOI and of the background signals at any frame within the tth block. Note that, the CSV coincides with the static model (7) when $T = 1$.

The practical meaning of the CSV model is illustrated in Fig. 1. While CSV admits that the SOI can change its position from block to block (the mixing vectors $\mathbf{a}_{k,t}$ depend on t), the block-independent separating vector $\mathbf{w}_k$ is sought such that extracts the speaker's voice from all positions visited during its movement. There are two main reasons for this: First, the achievable interference-to-signal ratio (ISR) depends on $\mathbf{w}_k$ so it has order $\mathcal{O}(N^{-1})$, compared to when $\mathbf{w}_k$ is block-dependent, which yields ISR of order $\mathcal{O}(N_b^{-1})$; this is confirmed by the theoretical study on Cramér-Rao bounds in [24]. Second, the CSV enables BSE methods to avoid the discontinuity problem mentioned in the previous section.

The CSV also brings a limitation. Formally, the mixture must obey the condition that for each k a separating vector exists such that $s_{k,t} = \mathbf{w}_k^H \mathbf{x}_{k,t}$ holds for every t; a condition that seems to be quite restrictive. Nevertheless, preliminary experiments in [23] have shown that this limitation is not crucial in practical situations and does not differ much from that of static methods (spatially overlapping speakers cannot be separated), especially when the number of microphones is high enough to provide sufficient degrees of freedom. When the speakers are static, the rule of thumb says that the speakers cannot be separated or, at least, are difficult to separate through spatial filtering, when their angular positions with respect to the microphone array are the same. Hence, moving speakers cannot be separated based on the CSV when their angular ranges with respect to the array during the recording are overlapping. The experimental part of this work presented in Section IV validates these findings.

2.3 Source model

In this section, we introduce the statistical model of the signals adopted from IVE. Samples (frames) of signals will be assumed to be identically and independently distributed (i.i.d.) within each block according to the probability density function (pdf) of the representing random variable.

Let $\mathbf{s}_t$ denote the vector component corresponding to the SOI, i.e., $\mathbf{s}_t = [s_{1,t}, \ldots, s_{K,t}]^T$. The elements of $\mathbf{s}_t$ are assumed to be uncorrelated (because they correspond to different frequency components of the SOI) but dependent, that is, their higher-order moments are taken into account [9]. Let $p_s(\mathbf{s}_t)$ denote the joint pdf of $\mathbf{s}_t$ and $p_{\mathbf{z}_{k,t}}(\mathbf{z}_{k,t})$ denote the pdf[2] of $\mathbf{z}_{k,t}$. For simplifying the notation, $p_s(\cdot)$ will be denoted without the index t although it is generally dependent on t. Since $\mathbf{s}_t$ and $\mathbf{z}_{1,t}, \ldots, \mathbf{z}_{K,t}$ are independent, their joint pdf within the tth block is equal to the product of marginal pdfs

$$p_s(\mathbf{s}_t) \cdot \prod_{k=1}^{K} p_{\mathbf{z}_{k,t}}(\mathbf{z}_{k,t}). \tag{11}$$

[2] We might consider a joint pdf of $\mathbf{z}_{1,t}, \ldots, \mathbf{z}_{K,t}$ that could possibly involve higher-order dependencies between the background components. However, since $p_{\mathbf{z}_{k,t}}(\cdot)$ is assumed Gaussian in this paper, and since signals from different mixtures (frequencies) are assumed to be uncorrelated, as in the standard IVA, we can directly consider $\mathbf{z}_{1,t}, \ldots, \mathbf{z}_{K,t}$ to be mutually independent.

By applying the transformation theorem to (11) using (10), from which it follows that

$$\begin{pmatrix} s_{k,t} \\ \mathbf{z}_{k,t} \end{pmatrix} = \mathbf{W}_{k,t}\mathbf{x}_{k,t} = \begin{pmatrix} \mathbf{w}_k^H\mathbf{x}_{k,t} \\ \mathbf{B}_{k,t}\mathbf{x}_{k,t} \end{pmatrix}, \tag{12}$$

the joint pdf of the observed signals from the tth block reads

$$p_{\mathbf{x}}(\{\mathbf{x}_{k,t}\}_k) = p_s\left(\left\{\mathbf{w}_k^H\mathbf{x}_{k,t}\right\}_k\right) \times \prod_{k=1}^{K} p_{\mathbf{z}_{k,t}}(\mathbf{B}_{k,t}\mathbf{x}_{k,t})|\det\mathbf{W}_{k,t}|^2. \tag{13}$$

Hence, the log-likelihood function as a function of the parameter vectors $\mathbf{w}_k$ and $\mathbf{a}_{k,t}$ and all available samples of the observed signals in the tth block is given by

$$\begin{aligned}\mathcal{L}(\{\mathbf{w}_k\}_k, \{\mathbf{a}_{k,t}\}_k|\{\mathbf{x}_{k,t}\}_k) \\ = \hat{\mathrm{E}}\left[\log p_s(\{\hat{s}_{k,t}\}_k)\right] + \sum_{k=1}^{K}\hat{\mathrm{E}}\left[\log p_{\mathbf{z}_{k,t}}(\hat{\mathbf{z}}_{k,t})\right] \\ + \log|\det\mathbf{W}_{k,t}|^2,\end{aligned} \tag{14}$$

where $\hat{s}_{k,t} = \mathbf{w}_k^H\mathbf{x}_{k,t}$ and $\hat{\mathbf{z}}_{k,t} = \mathbf{B}_{k,t}\mathbf{x}_{k,t}$ denote the current estimate of the SOI and of the background signals, respectively.

In BSS and BSE, the true pdfs of the original sources are not known, so suitable model densities have to be chosen in order to derive a contrast function based on (14). To find an appropriate surrogate of $p_s(\mathbf{s}_t)$, the variance of SOI, which can be changing from block to block[3] has to be taken into account. Let $f(\cdot)$ be a pdf corresponding to a normalized non-Gaussian random variable. To reflect the block-dependent variance, $p_s(\mathbf{s}_t)$ should be replaced by

$$p_s(\mathbf{s}_t) \approx f\left(\left\{\frac{s_{k,t}}{\sigma_{k,t}}\right\}_k\right)\left(\prod_{k=1}^{K}\sigma_{k,t}\right)^{-2}, \tag{15}$$

where $\sigma_{k,t}^2$ denotes the variance of $s_{k,t}$. Its unknown value is replaced by the sample-based variance of $\hat{s}_{k,t}$, which is equal to $\hat{\sigma}_{k,t} = \sqrt{\mathbf{w}_k^H\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}$ where $\widehat{\mathbf{C}}_{k,t} = \hat{\mathrm{E}}\left[\mathbf{x}_{k,t}\mathbf{x}_{k,t}^H\right]$ is the sample-based covariance matrix of $\mathbf{x}_{k,t}$.

It is worth noting that in the case of the static mixing model, i.e. when $T = 1$, it can be assumed that $\sigma_{k,t}^2 = 1$ because of the scaling ambiguity.

Similarly to [13], the pdf of the background is assumed to be circular Gaussian with zero mean and (unknown) covariance matrix $\mathbf{C}_{\mathbf{z}_{k,t}} = \mathrm{E}\left[\mathbf{z}_{k,t}\mathbf{z}_{k,t}^H\right]$, i.e., $p_{\mathbf{z}_{k,t}} \sim \mathcal{CN}(0, \mathbf{C}_{\mathbf{z}_{k,t}})$. Next, by Eq. (15) in [13] it follows that $|\det\mathbf{W}_{k,t}|^2 = |\gamma_{k,t}|^{2(d-2)}$, which corresponds to the third term in (14).

Now, by replacing the unknown pdfs in (14) and by neglecting the constant terms, we obtain the contrast function in the form

$$\begin{aligned}\mathcal{C}(\{\mathbf{w}_k\}_k, \{\mathbf{a}_{k,t}\}_{k,t}) \\ = \frac{1}{T}\sum_{t=1}^{T}\Bigg\{\hat{\mathrm{E}}\left[\log f\left(\left\{\frac{\mathbf{w}_k^H\mathbf{x}_{k,t}}{\hat{\sigma}_{k,t}}\right\}_k\right)\right] - \sum_{k=1}^{K}\log(\hat{\sigma}_{k,t})^2 \\ - \sum_{k=1}^{K}\hat{\mathrm{E}}\left[\mathbf{x}_{k,t}^H\mathbf{B}_{k,t}^H\mathbf{C}_{\mathbf{z}_{k,t}}^{-1}\mathbf{B}_{k,t}\mathbf{x}_{k,t}\right] \\ + (d-2)\sum_{k=1}^{K}\log|\gamma_{k,t}|^2\Bigg\}.\end{aligned} \tag{16}$$

The nuisance parameter $\mathbf{C}_{\mathbf{z}_{k,t}}$ will later be replaced by its sample-based estimate $\widehat{\mathbf{C}}_{\mathbf{z}_{k,t}} = \hat{\mathrm{E}}\left[\hat{\mathbf{z}}_{k,t}\hat{\mathbf{z}}_{k,t}^H\right]$.

3 Proposed algorithm

3.1 Orthogonal constraint

Finding the maximum of (16) subject to the separating and mixing vectors leads to their consistent estimation, hence to the solution of the BSE problem. The parameter vectors are linked through the distortionless constraint given by (6). However, as was already noticed in previous publications [13, 22, 25], this constraint appears to be too weak as it does not guarantee that both vectors finally found by an algorithm correspond to the SOI. Therefore, an additional constraint has to be imposed.

The orthogonal constraint (OGC) ensures that the current estimate of the SOI $\hat{s}_{k,t} = \mathbf{w}\mathbf{x}_{k,t}$ has zero sample correlation with the background signals $\hat{\mathbf{z}}_{k,t} = \mathbf{B}\mathbf{x}_{k,t}$. Hence the constraint is that $\hat{\mathrm{E}}\left[\hat{s}_{k,t}\hat{\mathbf{z}}_{k,t}^H\right] = \mathbf{w}_k^H\widehat{\mathbf{C}}_{k,t}\mathbf{B}_{k,t} = \mathbf{0}$, for every k and t, under the condition given by (6). In Appendix A in [13], it is shown that the OGC can be imposed by making $\mathbf{a}_{k,t}$ fully dependent on $\mathbf{w}_k$ through

$$\mathbf{a}_{k,t} = \frac{\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}{\mathbf{w}_k^H\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}, \quad t = 1, \ldots, T. \tag{17}$$

Alternatively, $\mathbf{w}_k$ can be considered as dependent on $\mathbf{a}_{k,t}$ [13]; however, we prefer the former formulation in this paper, because in the proposed algorithm, the optimization proceeds through the separating vectors $\mathbf{w}_k$.

3.2 Auxiliary function-based algorithm

In [20], N. Ono derived the AuxIVA algorithm using an auxiliary function-based optimization (AFO) technique. AuxIVA provides a much faster and more stable alternative to the natural gradient-based algorithm from [9]. The main principle of the AFO technique lies in replacing the first term in (16) by a majorizing term involving an auxiliary variable. The modified contrast function is named the auxiliary function. It is optimized in the auxiliary and normal variables alternately, by which the maximum of the original contrast function is found.

[3] The variance can be changing from block to block not only due to the signal nonstationarity, but also because of the movements of the source.

Very recently, a modification of AuxIVA for the blind extraction of q sources, where $q < d$, has been proposed in [22]; the algorithm is named OverIVA. In this section, we will apply the AFO technique to find the maximum of (16). The resulting algorithm, which could be seen as a special variant of OverIVA designed for $q = 1$ and as an extension for $T > 1$, will be called CSV-AuxIVE.

To find the suitable majorant of the first term of the contrast function (16) we can follow the original Theorem 1 from [20].

Theorem 1 *Let S_G be a set of real-valued functions of a vector variable $\mathbf{u}$ defined as*

$$S_G = \{G(\mathbf{u}) | G(\mathbf{u}) = G_R(\|\mathbf{u}\|_2)\}, \tag{18}$$

where $G_R(r)$ is a continuous and differentiable function of a real variable r satisfying that $\frac{G'_R(r)}{r}$ is continuous everywhere and is monotonically decreasing in $r \geq 0$. Then, for any $G(\mathbf{u}) = G_R(\|\mathbf{u}\|_2) \in S_G$,

$$G(\mathbf{u}) \leq \frac{G'_R(r_0)}{2r_0}\|\mathbf{u}\|_2^2 + \left(G_R(r_0) - \frac{r_0 G'_R(r_0)}{2}\right) \tag{19}$$

holds for any $\mathbf{u}$ and $r_0 \geq 0$. The equality holds if and only if $r_0 = \|\mathbf{u}\|_2$.

Proof See [20]. □

Now, let $G(\mathbf{u}) = \log f(\mathbf{u})$ and assume that the conditions of the theorem are satisfied. Then, by applying Theorem 1 on the tth block of the first term of (16) we get a relation

$$\begin{aligned}\hat{\mathrm{E}}\left[\log f\left(\left\{\frac{s_{k,t}}{\sigma_{k,t}}\right\}_k\right)\right] &\leq \hat{\mathrm{E}}\left[\frac{G'_R(r_t)}{2r_t}\sum_k^K \left|\frac{s_{k,t}}{\hat{\sigma}_{k,t}}\right|^2\right] + R_t\\ &= \hat{\mathrm{E}}\left[\frac{G'_R(r_t)}{2r_t}\sum_k^K \frac{\mathbf{w}_k^H\mathbf{x}_{k,t}\mathbf{x}_{k,t}^H\mathbf{w}_k}{\hat{\sigma}_{k,t}^2}\right] + R_t\\ &= \sum_k^K \frac{1}{2}\frac{1}{\hat{\sigma}_{k,t}^2}\mathbf{w}_k^H \hat{\mathrm{E}}\left[\frac{G'_R(r_t)}{r_t}\mathbf{x}_{k,t}\mathbf{x}_{k,t}^H\right]\mathbf{w}_k + R_t\end{aligned} \tag{20}$$

where r_t is an auxiliary variable and R_t depends purely on r_t; the equality holds if and only if $r_t = \sqrt{\sum_{k=1}^K |\mathbf{w}_k^H\mathbf{x}_{k,t}|^2/\hat{\sigma}_{k,t}^2}$. By applying (20) in (16), the auxiliary function obtains a form

$$\begin{aligned}&Q\left(\{\mathbf{w}_k, \mathbf{a}_{k,t}, r_t\}_{k,t}\right)\\ &= \frac{1}{T}\sum_{t=1}^T\sum_{k=1}^K\left\{\frac{1}{2}\frac{\mathbf{w}_k^H\mathbf{V}_{k,t}\mathbf{w}_k}{\hat{\sigma}_{k,t}^2} - \log\hat{\sigma}_{k,t}^2\right.\\ &\quad\left. - \hat{\mathrm{E}}\left[\hat{\mathbf{z}}_{k,t}^H\mathbf{C}_{\mathbf{z}_{k,t}}^{-1}\hat{\mathbf{z}}_{k,t}\right] + (d-2)\log|\gamma_{k,t}|^2\right\} + R_t,\end{aligned} \tag{21}$$

where

$$\mathbf{V}_{k,t} = \hat{\mathrm{E}}[\,\varphi(r_t)\mathbf{x}_{k,t}\mathbf{x}_{k,t}^H]\,, \tag{22}$$

and $\varphi(r) = \frac{G'_R(r)}{r}$. Now, we can see that

$$\mathcal{C}(\{\mathbf{w}_k, \mathbf{a}_{k,t}\}_{k,t}) \leq Q(\{\mathbf{w}_k, \mathbf{a}_{k,t}, r_t\}_{k,t}), \tag{23}$$

where both sides are equal if and only if $r_t = \sqrt{\sum_{k=1}^K \left|\mathbf{w}_k^H\mathbf{x}_{k,t}\right|^2/\hat{\sigma}_{k,t}^2}$ for every $t = 1, \ldots, T$, so (21) is a valid auxiliary function.

The optimization of Q proceeds alternately in the auxiliary variables r_t and the normal variables $\mathbf{w}_k$. The optimum of (21) in the auxiliary variables is obtained simply by putting $r_t = \sqrt{\sum_{k=1}^K \left|\mathbf{w}_k^H\mathbf{x}_{k,t}\right|^2/\hat{\sigma}_{k,t}^2}$ into (22). To find the minimum in the normal variables, the partial derivative of the auxiliary function (21) is taken with respect to $\mathbf{w}_k$ when r_t is independent, and $\mathbf{a}_{k,t}$ are dependent through the OGC (17). The derivative is put equal to zero, which forms equations for the new update of the separating vectors.

For the derivative of the first and second term in (21), the following identities are used, which come from straightforward computations using the Wirtinger calculus [26] and by using the OGC (17):

$$\frac{\partial}{\partial\mathbf{w}_k^H}\frac{1}{\hat{\sigma}_{k,t}^2} = -\frac{\mathbf{a}_{k,t}}{\hat{\sigma}_{k,t}^2}, \tag{24}$$

$$\frac{\partial}{\partial\mathbf{w}_k^H}\log\hat{\sigma}_{k,t}^2 = \mathbf{a}_{k,t}. \tag{25}$$

The computation of the derivative of the third and fourth term of (21) is lengthy due to the dependence of the parameters through the OGC constraint. To simplify, we can use Equation 33 and Appendix C in [13], where the derivative is actually computed for the case $K = 1$ and $T = 1$, from which it follows that the result is equal to $\sum_{k=1}^K \mathbf{a}_{k,t}$. By putting the derivatives of all the term together, we obtain

$$\begin{aligned}&\left.\frac{\partial Q\left(\{\mathbf{w}_k, \mathbf{a}_{k,t}, r_t\}_{k,t}\right)}{\partial\mathbf{w}_k^H}\right|_{\text{w.r.t. (17)}}\\ &= \frac{1}{T}\sum_{t=1}^T\left\{\frac{\mathbf{V}_{k,t}\mathbf{w}_k}{2\hat{\sigma}_{k,t}^2} - \frac{\mathbf{w}_k^H\mathbf{V}_{k,t}\mathbf{w}_k}{2\hat{\sigma}_{k,t}^2}\mathbf{a}_{k,t} - \mathbf{a}_{k,t} + \mathbf{a}_{k,t}\right\}.\end{aligned} \tag{26}$$

The close-form solution of the equation when (26) is put equal to zero cannot be derived in general. Our proposal is to take

$$\mathbf{w}_k = \left(\sum_{t=1}^T \frac{\mathbf{V}_{k,t}}{\hat{\sigma}_{k,t}^2}\right)^{-1}\sum_{t=1}^T \frac{\mathbf{w}_k^H\mathbf{V}_{k,t}\mathbf{w}_k}{\hat{\sigma}_{k,t}^2}\mathbf{a}_{k,t}, \tag{27}$$

which is the solution of a linearized equation where the terms $\mathbf{w}_k^H\mathbf{V}_{k,t}\mathbf{w}_k$ and $\hat{\sigma}_{k,t}^2$ are treated as constants that are independent of $\mathbf{w}_k$. Hence, the general update rules of CSV-AuxIVE are as follows:

$$r_t = \sqrt{\sum_{k=1}^{K} \left|\mathbf{w}_k^H \mathbf{x}_{k,t}\right|^2 / \hat{\sigma}_{k,t}^2}, \tag{28}$$

$$\mathbf{V}_{k,t} = \hat{\mathrm{E}}\left[\varphi(r_t)\mathbf{x}_{k,t}\mathbf{x}_{k,t}^H\right], \tag{29}$$

$$\mathbf{a}_{k,t} = \frac{\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}{\mathbf{w}_k^H \widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}, \tag{30}$$

$$\hat{\sigma}_{k,t} = \sqrt{\mathbf{w}_k^H \widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}, \tag{31}$$

$$\mathbf{w}_k = \left(\sum_{t=1}^{T} \frac{\mathbf{V}_{k,t}}{\hat{\sigma}_{k,t}^2}\right)^{-1} \sum_{t=1}^{T} \frac{\mathbf{w}_k^H \mathbf{V}_{k,t}\mathbf{w}_k}{\hat{\sigma}_{k,t}^2}\mathbf{a}_{k,t}, \tag{32}$$

$$\mathbf{w}_k \leftarrow \mathbf{w}_k / \sqrt{\sum_{t=1}^{T} \mathbf{w}_k^H \mathbf{V}_{k,t}\mathbf{w}_k}. \tag{33}$$

The last step, which performs a normalization of the updated separating vectors, has been found important to the stability of the convergence. After the convergence is achieved, the separating vectors are re-scaled using least squares to reconstruct the images of the SOI on a reference microphone [27].

In our implementation, we consider the standard nonlinearity $\varphi(r_t) = r_t^{-1}$ proposed in [20], which is known to be suitable for super-Gaussian signals such as speech. For this particular choice, we propose one more modification in the proposed algorithm: compared to (28), r_t is put equal to $\sqrt{\sum_{k=1}^{K} \left|\mathbf{w}_k^H \mathbf{x}_{k,t}\right|^2}$. We have experienced improved convergence speed with this modification. The pseudo-code is summarized in Algorithm 1,

Algorithm 1: Pseudo-code of CSV-AuxIVE

```
Input: $\mathbf{x}_{k,t}, \mathbf{w}_k^{\mathrm{ini}}\ (k, t = 1, 2, \dots)$, NumIter
Output: $\mathbf{a}_{k,t}, \mathbf{w}_k$
foreach $k = 1, \dots, K,\ t = 1, \dots, T$ do
    $\widehat{\mathbf{C}}_{k,t} = \hat{\mathrm{E}}\left[\mathbf{x}_{k,t}\mathbf{x}_{k,t}^H\right]$;
    $\mathbf{w}_k = \mathbf{w}_k^{\mathrm{ini}} / \left(\mathbf{w}_k^{\mathrm{ini}}\right)_1$;
end
Iter = 0;
repeat
    foreach $t = 1 \dots T$ do
        foreach $k = 1 \dots K$ do
            $\hat{\sigma}_{k,t} \leftarrow \sqrt{\mathbf{w}_k^H \widehat{\mathbf{C}}_{k,t}\mathbf{w}_k}$;
        end
        $r_t \leftarrow \sqrt{\sum_{k=1}^{K} \left|\mathbf{w}_k^H \mathbf{x}_{k,t}\right|^2}$;
        foreach $k = 1 \dots K$ do
            $\mathbf{a}_{k,t} \leftarrow \left(\mathbf{w}_k^H \widehat{\mathbf{C}}_{k,t}\mathbf{w}_k\right)^{-1} \left(\widehat{\mathbf{C}}_{k,t}\mathbf{w}_k\right)$;
            $\mathbf{V}_{k,t} \leftarrow \hat{\mathrm{E}}\left[\frac{1}{r_t}\mathbf{x}_{k,t}\mathbf{x}_{k,t}^H\right]$;
        end
    end
    foreach $k = 1 \dots K$ do
        Compute $\mathbf{w}_k^H$ according (32);
        $\mathbf{w}_k \leftarrow \mathbf{w}_k / \sqrt{\sum_{t=1}^{T} \mathbf{w}_k^H \mathbf{V}_{k,t}\mathbf{w}_k}$;
    end
    Iter ← Iter + 1;
until Iter < NumIter;
```

3.3 Semi-supervised CSV-AuxIVE

Owing to the indeterminacy of order in BSE it is not, in general, known which source is currently being extracted. The crucial problem is to ensure that the signal being extracted actually corresponds to the desired SOI. In BOGIVE$_{\mathbf{w}}$ as well as in CSV-AuxIVE, this can be influenced only through the initialization. The question of convergence of the BSE algorithms has been considered in [13].

Several approaches ensuring the global convergence have been proposed, most of which are based on additional constraints assuming prior knowledge, e.g., about the source position or a reference signal [18, 28–30]. Recently, an unconstrained supervised IVA using so-called pilot signals has been proposed in [31]. The pilot signal, which is assumed to be available as prior information, is a signal that is mutually dependent with the corresponding source signal. Therefore, the pilot signal and the frequency components of the source have a joint pdf. In the piloted IVA, the pilot signals are used as constant "frequency components" in the joint pdf model, which is helpful in solving the permutation problem as well as the ambiguous order of the separated sources. In [13], the idea has been applied in IVE, where the pilot signal related to the SOI is assumed to be available.

Let the pilot signal (dependent on the SOI and independent of the background) be represented on the tth block by o_t (o_t is denoted without index k; nevertheless, it can also be k-dependent). Let the joint pdf of $\mathbf{s}_t$ and o_t be $p(\mathbf{s}_t, o_t)$. Then, similarly to (13), the pdf of the observed data within the tth block is given by

$$p_{\mathbf{x}}(\{\mathbf{x}_k\}_{k,t}) = p\left(\left\{\mathbf{w}_k^H \mathbf{x}_{k,t}\right\}_{k,t}, o_t\right) \times \prod_{k=1}^{K} p_{\mathbf{z}_{k,t}}(\mathbf{B}_{k,t}\mathbf{x}_{k,t}) |\det \mathbf{W}_{k,t}|^2. \tag{34}$$

Comparing this expression with (13) and taking into account the fact that o_t is independent of the mixing model parameters, it can be seen that the modification of CSV-AuxIVE towards the use of pilot signals is straightforward.

In particular, provided that the model pdf $f\left(\left\{\mathbf{w}_k^H\mathbf{x}_k\right\}_{k,t}, o_t\right)$ replacing the unknown $p(\cdot)$ meets the conditions of Theorem 1, the piloted algorithm has exactly the same steps as the non-piloted one with a sole difference that the non-linearity $\varphi(\cdot)$ also depends on o_t. Therefore, the Eq. 28 will have form

$$r_t = \sqrt{\sum_{k=1}^{K} \left|\mathbf{w}_k^H \mathbf{x}_{k,t}\right|^2 + \eta^2 |o_t|^2}, \tag{35}$$

for $t = 1, \ldots, T$, where η is a hyperparameter controlling the influence of the pilot signal [31].

Consequently, the semi-supervised of CSV-AuxIVE, in this manuscript referred as piloted CSV-AuxIVE, is obtained by replacing the update step (28) with (35).

Finding a suitable pilot signal poses an application-dependent problem. For example, outputs of voice activity detectors were used to pilot the separation of simultaneously talking people in [31]. Similarly, a video-based lip-movement detection was considered in [32]. A video-independent solution was proposed in [33] using spatial information about the area in which the speaker is located. Recently, the approach utilizing speaker identification was proposed in [34] and further improved in [35]. All of these approaches have been shown to be very useful, even though the used pilot signals contain residual noise and interference. The design of a pilot signal is a topic beyond the scope of this paper. Therefore, in the experimental part of this paper, we consider only oracle pilots as proof of concept.

4 Experimental validation

In this section, we present results of experiments with simulated mixtures as well as real-world recordings of moving speakers. Our goal is to show the usefulness of the CSV mixing model and to compare the performance characteristics of the proposed algorithm with other state-of-the-art methods.

4.1 Simulated room

In this example, we inspect spatial characteristics of de-mixing filters obtained by the blind algorithms when extracting a moving speaker in a room simulated by the image method [36].

4.1.1 Experimental setup

The room has dimensions 4 × 4 × 2.5 (width×*length*× height) meters and $T_{60} = 100$ ms. A linear array of five omnidirectional microphones is located so that its center is at the position $(1.8, 2, 1)$ m, and the array axis is parallel with the room width. The spacing between microphones is 5 cm.

The target signal is a 10 s long female utterance from TIMIT dataset [37]. During speech, the speaker is moving at a constant speed on a 38° arc at a one-meter distance from the center of the array; the situation is illustrated in Fig. 2a. The starting and ending positions are $(1.8, 3, 1)$ m and $(1.2, 2.78, 1)$ m, respectively. The movement is simulated by 20 equidistantly spaced RIRs on the path, which correspond to half-second intervals of speech, whose overlap was smoothed by windowing. As an interferer, a point source emitting white Gaussian noise is located at the position $(2.8, 2, 1)$ m; that is, at a 1-m distance to the right from the array.

The mixture of speech and noise has been processed in order to extract the speech signal by the following methods: OGIVE$_\mathrm{w}$ [13], BOGIVE$_\mathrm{w}$ (the extension of OGIVE$_\mathrm{w}$ allowing for the CSV; derived in the supplementary material of this article), OverIVA with $m = 1$ [22], which corresponds with CSV-AuxIVE when $T = 1$, and CSV-AuxIVE. All methods operate in the STFT domain with the FFT length of 512 samples and 128 samples hop-size; the sampling frequency is $f_s = 16$ kHz. Each method has been initialized by the direction of arrival of the desired speaker signal at the beginning of the sequence. The other parameters of the methods are listed in Table 1.

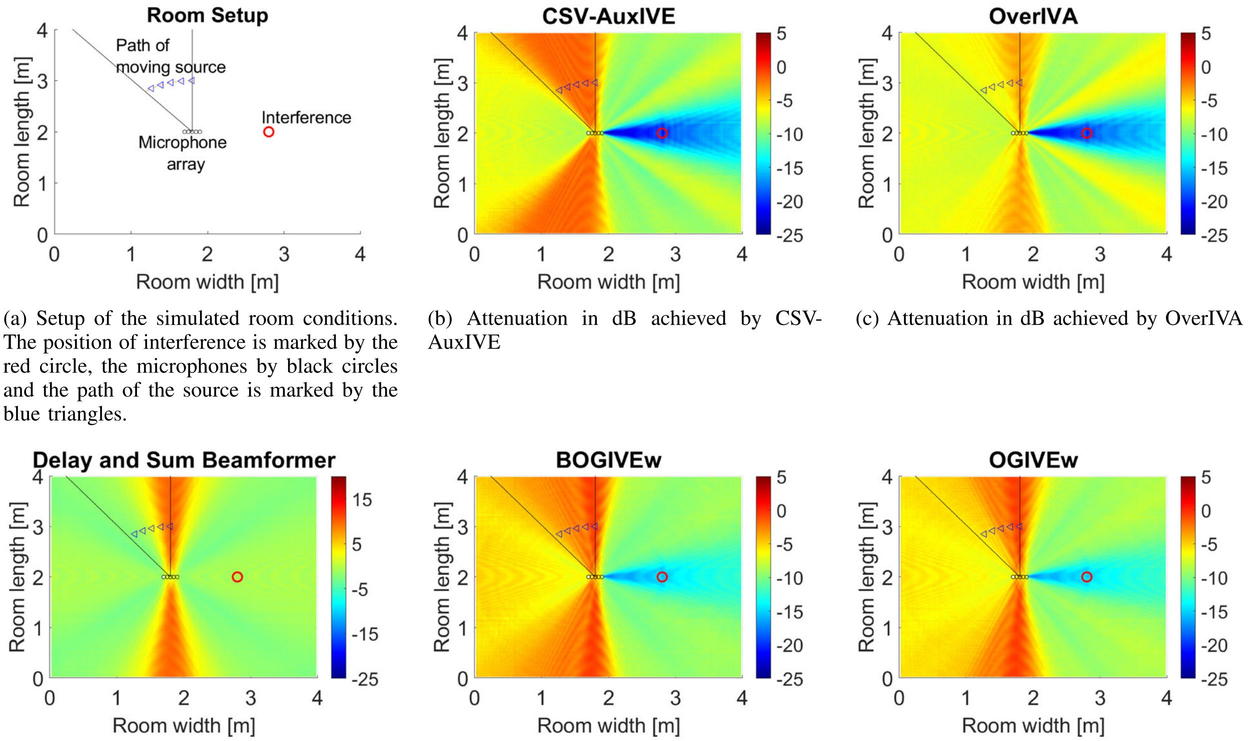
In order to visualize the performance of the extracting filters, a 2 × 2 cm-spaced regular grid of positions spanning the whole room is considered. Microphone responses (images) of a white Gaussian noise signal emitted from each position on the grid have been simulated. The extracting filter of a given algorithm is applied to the microphone responses, and the output power is measured. The average ratio between the output power and the power of the input signals reflects the attenuation of the white noise signal originating from the given position.

4.1.2 Results

The attenuation maps of the compared methods are shown in Fig. 2b through 2f, and Table 2 shows the attenuation for specific points in the room. In particular, the first five columns in the table correspond to the speaker's positions on the movement path at angles 0° through 32°. The last column corresponds to the position of the interferer.

Figure 2d shows the map of the initial filter corresponding to the delay-and-sum (D&S) beamformer steered towards the initial position of the speaker. The beamformer yields a gentle gain in the initial direction with no attenuation in the direction of the interferer.

The compared blind methods steer a spatial null towards the interferer and try to pass through the target signal. However, OverIVA and OGIVE$_\mathrm{w}$ tend to pass through only a narrow angular range (probably the most significant part of the speech). By contrast, the spatial beam steered by CSV-AuxIVE towards the speaker spans the whole angular range where the speaker has



(a) Setup of the simulated room conditions. The position of interference is marked by the red circle, the microphones by black circles and the path of the source is marked by the blue triangles.

(b) Attenuation in dB achieved by CSV-AuxIVE

(c) Attenuation in dB achieved by OverIVA

(d) Attenuation in dB achieved by Delay and sum Beamformer

(e) Attenuation in dB achieved by BOGIVE$_w$

(f) Attenuation in dB achieved by OGIVE$_w$

Fig. 2 Setup of the simulated room and the attenuation in dB achieved by DOA, OGIVE$_w$, BOGIVE$_w$, OverIVA, and CSV-AuxIVE from the experiment in Section 4.1

appeared during the movement. BOGIVE$_w$ performs similarly, however, its performance is poorer, perhaps due to its slower convergence or proneness to getting stuck in a local extreme. The convergence comparison of BOGIVE$_w$ and CSV-AuxIVE is shown in Fig. 3. The nulls steered towards the interferer by OverIVA and CSV-AuxIVE are more attenuating compared to the gradient methods. In conclusion, these results confirm the ability of the blind algorithms to extract the moving source gained through of the CSV mixing model. The results also show better convergence properties of CSV-AuxIVE over BOGIVE$_w$.

4.2 Moving speakers simulated by wireless loudspeaker attached to turning arm

The goal of this experiment is to compare the performance of algorithms as they depend on the range and speed of movements of the sources.

4.2.1 *Experimental setup*

We have recorded a dataset of speech utterances that were played from a wireless loudspeaker (JBL GO 2) attached to a manually actuated rotating arm. The length of each utterance is 31 s. Sounds were recorded with 16 kHz sampling rate using a linear array of four microphones with 16 cm spacing. The array center was placed at the arm's pivot. This allows the apparatus to simulate circular movements of sources at a radius of approx. 1 m. The recording setup was placed in an open-space 12 x 8 x 2.6 m room with a reverberation time $T_{60} \approx$ 500ms. The recording setup is shown in Fig. 4.

Table 1 Parameter setup for the tested methods in the simulated room

Method	# Iterations	Step size μ	Block size N_b
OGIVE$_w$	1000	0.2	n/a
BOGIVE$_w$	1000	0.2	250 frames
OverIVA	100	n/a	n/a
CSV-AuxIVE	100	n/a	250 frames

Table 2 The attenuation (dB) in selected points on the source path and in the position of the interferer

	0°	8°	16°	24°	32°	Interferer
OGIVE$_w$	**−1.09**	**−1.36**	−2.02	−4.56	−5.08	−15.81
BOGIVE$_w$	−1.2	−2.14	−1.69	−3.12	−3.87	−15.86
OverIVA	−5.85	−3.99	−3.08	−4.39	−5.12	**−23.73**
CSV-AuxIVE	−3.22	−1.74	**−1.27**	**−2.09**	**−2.67**	−18.51

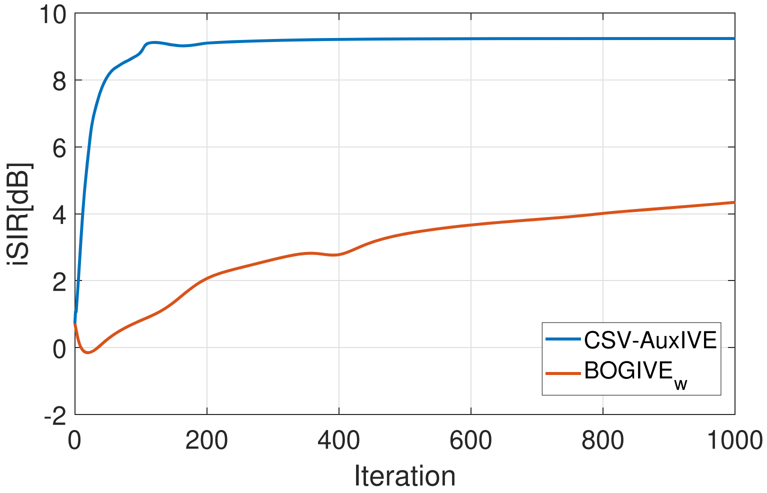


Fig. 3 The convergence of CSV-AuxIVE and BOGIVE$_w$ in term of SIR improvement averaged over 200 random mixtures in the experiment of Section 4.3; $T_{60} = 100ms$

The dataset consists of two individual, spatially separated sources. The SOI is represented by a male speech utterance and is confined to the angular interval from 0° through 90°. The interference (IR) is represented by a female speech utterance and is confined to the interval of −90° through 0°. The list of recordings is described in Table 3. The recordings along with videos of the recording process are available online (see links at the end of this article).

Thirty-six mixtures were created by combining the SOI and IR recordings in Table 3; the input SIR was set to 10 dB. The following three algorithms were compared: CSV-AuxIVE with the length of blocks set to 100 frames, the original AuxIVA algorithm [20], and a sequential on-line variant of AuxIVA (On-line AuxIVA) from [17] with the time-window length of 20 frames and the forgetting factor set to 0.95. The algorithms operated in the STFT domain with 1024 samples per frame and 768 samples overlap. The off-line algorithms were stopped after 100 iterations. In case of AuxIVA and On-line AuxIVA, the output channel containing the SOI was determined based on the output SDR.

Performance was evaluated using segmental measures: normalized SIR (nSIR), SDR improvement (iSDR), and the average SOI attenuation (Attenuation); nSIR is the ratio of the powers of the SOI and IR in the extracted signal where each segment is normalized to unit variance; SDR is computed using BSS_eval [38]. While iSDR and Attenuation reflect the loss of power of the SOI in the extracted signal, nSIR reflects also the IR cancelation. The length of segments was set to 1 s.

4.2.2 Results

The results in Fig. 5 show that AuxIVA and On-line AuxIVA perform well only when the SOI is static. Their performances drop when the SOI moves. On-line AuxIVA

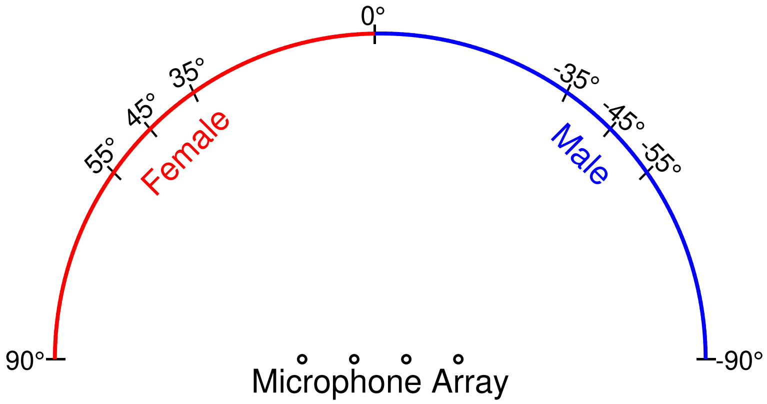


Fig. 4 The recording setup for the experiment in Section 4.2. Black circles denote the position of microphones. The red and blue lines show the path of the female and male speakers, respectively

Table 3 Angular intervals and speed of SOI and IR movements in experiment 4.2

Recording index	1	2	3	4	5	6
Movement speed	**Static**	**Slow**	**Fast**	**Slow**	**Fast**	**Very fast**
Range of the movement: starting angle, ending angle						
	Static	Small	Small	Large	Large	Large
SOI (male)	− 45°	− 55°, − 35°	− 55°, − 35°	− 90° ,0°	− 90°, 0°	− 90°, 0°
IR (female)	45°	35°, 55°	35°, 55°	0°, 90°	0°, 90°	0°, 90°

is slightly less sensitive to the SOI movement compared to AuxIVA due to its adaptability. However, the overall performance of On-line AuxIVA is low, because the algorithm works with limited context.

CSV-AuxIVE shows significantly smaller sensitivity to the SOI movements than the compared algorithms. This is mainly reflected by Attenuation, which is only slightly growing with the increasing range and speed of the SOI movement. The higher performance of CSV-AuxIVE in terms of iSDR and nSIR compared to AuxIVA and On-line AuxIVA confirms the new ability of the proposed algorithm gained due to the CSV mixing model.

The IR movements cause the performance of AuxIVA and CSV-AuxIVE to decrease with the growing range of the IR movement (small and large). The speed of movement seems to play a minor role. This can be explained by the fact that the off-line algorithms estimate time-invariant spatial filters which project two distinct beams: one towards the entire angular area occupied by the SOI and one towards the area occupied by the IR. The former beam should pass the incoming signal through while the latter beam should attenuate it. Provided that the estimated filters satisfy these requirements, as long as the sources stay within their respective beams, the speed with

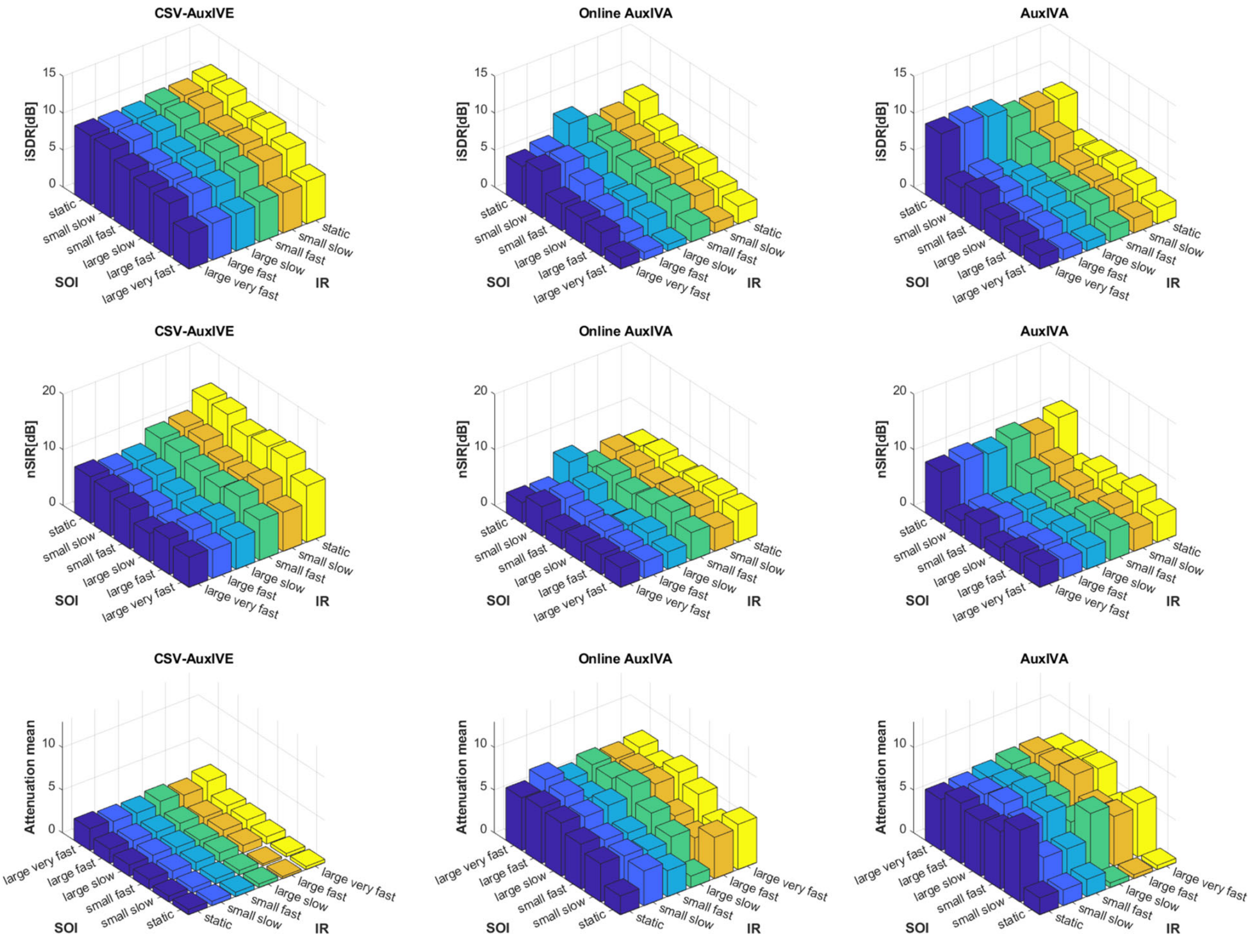


Fig. 5 The accuracy of the blind extraction of the SOI in terms of iSDR, nSIR, and Attenuation in the experiment in Section 4.2. The indices on the SOI and IR axes correspond with Table 3. Please note that, for better readability, the axes of the plots showing the Attenuation are *reversed*

which they move does not matter. For the estimation of the filters based on the CSV itself, the speakers should be approximately static within each block as the mixing vectors are assumed constant within the blocks. Hence, the allowed speed should not be too high compared to the block length.

In conclusion, the results reflect the theoretical capabilities of the algorithms, or, more specifically, of the filters that they can estimate. AuxIVA can steer only a narrow beam towards the SOI, which can therefore be extracted efficiently only if the SOI is not moving. On-line AuxIVA can steer a narrow beams in the adaptive way, however, the accuracy is lower due to a small context of data. CSV-AuxIVE can reliably extract the SOI from a wider area within the entire context of the data.

4.3 Real-world scenario using the MIRaGe database

This experiment is designed to provide an exhaustive test of the compared methods in challenging noisy situations where the target speaker is performing small movements within a confined area.

4.3.1 Experimental setup

Recordings are simulated using real-world room impulse responses (RIRs) taken from the MIRaGe database [39]. MIRaGe provides measured RIRs between microphones and a source whose possible positions form a dense grid within a 46 × 36 × 32 cm volume. MIRaGe is thus suitable for our experiment, as it enables us to simulate small speaker movements in a real environment.

The database setup is situated in an acoustic laboratory which is a 6 × 6 × 2.4 m rectangular room with variable reverberation time. Three reverberation levels with T_{60} equal to 100, 300, and 600 ms are provided. The speaker's area involves 4104 positions which form the cube-shaped grid with spacings of 2-by-2 cm over the x and y axes and 4 cm over the z axis. MIRaGe also contains a complementary set of measurements that provide information about the positions placed around the room perimeter with spacing of approx. 1 m, at a distance of 1 m from the wall. These positions are referred to as the out-of-grid positions (OOG). All measurements were recorded by six static linear microphone arrays (5 mics per array with the inter-microphone spacing of -13, -5, 0, $+5$, and $+13$ cm relative to the central microphone); for more details about the database, see [39].

In the present experiment, we use Array 1, which is at a distance of 1 m from the center of the grid, and the T_{60} settings of 100 and 300 ms. For each setting, 3840 noisy observations of a moving speaker were synthesized as follows: each mixture consists of a moving SOI, one static interfering speaker and noise. The SOI is moving randomly over the grid positions. The movement is simulated so that the position is changed every second. The new position is randomly selected from all positions whose maximum distance from the current position is 4 in both the x and y axes. The transition between positions is smoothed using the Hamming window of a length of $f_s/16$ with one-half overlaps. The interferer is located in a random OOG position between 13 through 24, while the noise signal is equal to a sum of signals that are located in the remaining OOG positions (out of 13 through 24).

As the SOI and interferer signal, clean utterances of 4 male and 4 female speakers from the CHiME-4 [40] dataset were selected; there are 20 different utterances, each having 10 s in length per speaker. The noise signals correspond to random parts of the CHiME-4 cafeteria noise recording. The signals are convolved with the RIRs to match the desired positions, and the obtained spatial images of the signals on microphones are summed up so that the interferer/noise ratio, as well as the ratio between the SOI and interference plus noise, is 0 dB.

The methods considered in the previous sections are compared. All these methods operate in the STFT domain with an FFT length of 1024 and a hop-size of 256; the sampling frequency is 16 kHz. The number of iterations is set to 150 and 2,000 for the offline AFO-based and the gradient-based methods, respectively. For the online AuxIVA, the number of iterations is set to 3 on each block. The block length in CSV-AuxIVE and BOGIVE$_{\mathrm{w}}$ is set to 150 frames. The online AuxIVA operates on block length of 50 frames with 75% overlap. The step-length in OGIVE$_{\mathrm{w}}$ and BOGIVE$_{\mathrm{w}}$ is set to $\mu = 0.2$. The initial separating vector corresponds to the D&S beamformer steered in front of the microphone array. As a proof of concept for the approaches discussed in Section 3.3, we also compare the piloted variants of OverIVA and CSV-AuxIVE where the pilot signal corresponds to the energy of ground truth SOI on the frames.

4.3.2 Results

The SOI is blindly extracted from each mixture for the IVE methods. For the IVA methods, the output channel was determined by output SIR. The result is evaluated through the improvement of the signal-to-interference-and-noise ratio (iSINR) and signal-to-distortion ratio (iSDR) defined as in [41] (SDR is computed after compensating for the global delay). The averaged values of the criteria are summarized in Table 4 together with the average time to process one mixture. For a deeper understanding to the results, we also analyze the histograms of iSINR by OverIVA and CSV-AuxIVE shown in Fig. 6.

Figure 6a shows the histograms over the full dataset of mixtures, while Fig. 6b is evaluated on a subset of mixtures in which the SOI has not moved away from the starting position by more than 5 cm; there are 288 mixtures of this kind. Now, we can observe two phenomena. First, it can be seen that OverIVA yields more results below 10 dB in

Table 4 The SINR improvement with standard deviation, SDR improvement with standard deviation and extraction fail percentage for the MIRaGe database experiment

	T60 100 ms			**T60 300ms**			**Average**
	Mean iSINR [dB]	**Mean iSDR [dB]**	**iSINR <−5 dB [%]**	**Mean iSINR [dB]**	**Mean iSDR [dB]**	**iSINR <−5 dB [%]**	**time per mixture [s]**
AuxIVA	7.23 ± 7.78	4.56 ± 2.15	**0**	5.44 ± 6.43	4.13 ± 1.88	**0**	11.37
AuxIVA online	5.35 ± 6.73	3.96 ± 3.01	**0**	4.12 ± 5.65	3.12 ± 1.98	**0**	16.32
OverIVA	7.55 ± 8.33	3.96 ± 2.14	8.83	5.34 ± 7.01	3.82 ± 2.00	8.43	**8.00**
CSV-AuxIVE	9.45 ± 7.24	4.02 ± 1.27	6.72	6.84 ± 6.52	3.48 ± 1.17	6.71	9.14
Piloted OverIVA	11.99 ± 5.42	5.10 ± 3.37	0.65	9.67 ± 4.58	3.00 ± 2.55	0.26	8.16
Piloted CSV-AuxIVE	**13.72 ± 3.51**	**6.14 ± 2.13**	**0**	**11.41 ± 3.54**	**4.73 ± 1.91**	0	9.14
BOGIVE_w	4.32 ± 5.15	3.14 ± 1.56	15.32	2.28 ± 3.15	1.98 ± 1.02	22.15	86.45
OGIVE_w	3.85 ± 4.33	3.58 ± 1.98	22.10	1.01 ± 2.17	2.14 ± 1.45	12.23	73.15

Fig. 6a than in Fig. 6b. This confirms that OverIVA performs better for the subset of mixtures where the SOI is almost static. The performance of CSV-AuxIVE tends to be rather similar for the full set and the subset. CSV-AuxIVE thus yields a more stable performance than the static model-based OverIVA when the SOI performs small movements. Second, the piloted methods yield iSINR <-5 dB in a much lower number of trials than the non-piloted methods, as confirmed by the additional criterion in Table 4. This shows that the piloted algorithms have significantly improved global convergence. Note that IVA algorithms achieved iSINR <-5 dB in 0% of cases. For the IVE algorithms, the percentage of iSINR <-5 dB reflects the rate of extractions of a different source. In contrast, for IVA algorithms, the sources are either successfully separated or not, e.g. iSINR is around 0 dB.

4.4 Speech enhancement/recognition on CHiME-4 datasets

We have verified the proposed methods using the noisy speech recognition task defined within the CHiME-4 challenge, specifically, the six-channel track [40].

4.4.1 Experimental setup

This dataset contains simulated (SIMU) and real-world[4] (REAL) utterances of speakers in multi-source noisy environments. The recording device is a tablet with six microphones, which is held by a speaker. Since some recordings involve microphone failures, the method from [42] is used to detect these failures. If detected, the malfunctioning channels are excluded from further processing of the given recording.

The experiment is evaluated in terms of word error rate (WER) as follows: the compared methods are used to extract speech from the noisy recordings. Then, the enhanced signals are forwarded to the baseline speech recognizer from [40]. The WER achieved by the proposed methods is compared with the results obtained on unprocessed input signals (Channel 5) and with the techniques listed below.

BeamformIt [43] is a front-end algorithm used within the CHiME-4 baseline system. It is a weighted delay-and-sum beamformer requiring two passes over the processed recording in order to optimize its inner parameters. We compare the original implementation of the technique available at [44].

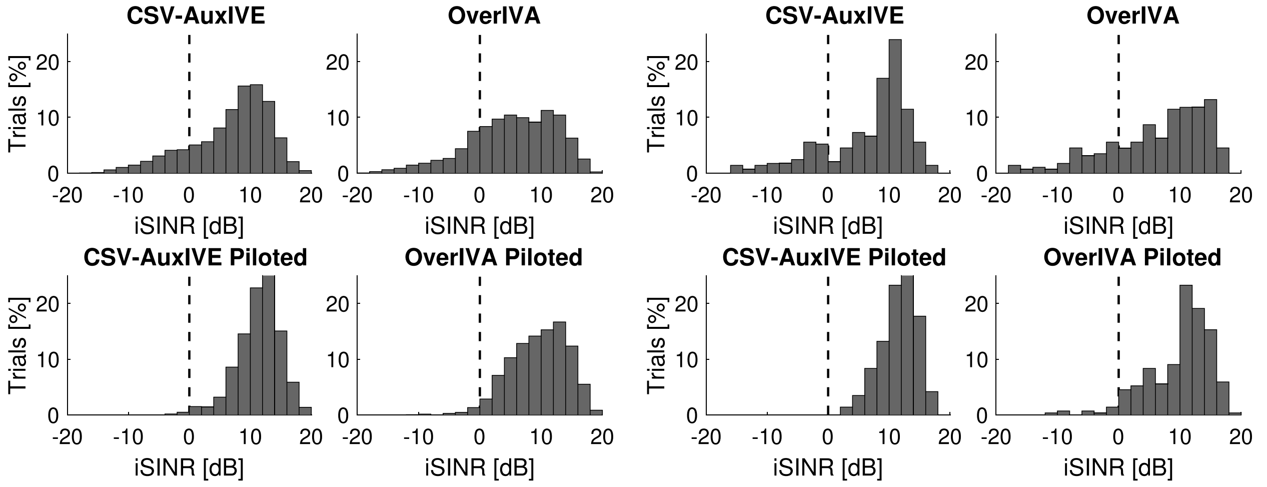
The generalized eigenvalue beamformer (GEV) is a front-end solution proposed in [45, 46]. It represents the most successful enhancers for CHiME-4 that rely on deep networks trained for the CHiME-4 data. In the implementation used here, a re-trained voice-activity-detector (VAD) is used where the training procedure was kindly provided by the authors of [45]. We utilize the feed-forward topology of the VAD and train the network using the training part of the CHiME-4 data. GEV utilizes the blind analytic normalization (BAN) postfilter to obtain its final enhanced output signal.

All systems/algorithms operate in the STFT domain with an FFT length of 512, a hop-size of 128 and use the Hamming window; the sampling frequency is 16 kHz. $\text{BOGIVE}_\mathbf{w}$ and CSV-AuxIVE are applied with $N_b = 250$, which corresponds to the block length of 2 s. This value has been selected to optimize the performance of these methods. All of the proposed methods are initialized by the relative transfer function (RTF) estimator from [47]; Channel 5 of the data is selected as the target (the spatial image of the speech signal of this channel is being estimated).

4.4.2 Results

The results shown in Table 5 indicate that all methods are able to improve the WER compared to the unprocessed case. The BSE-based methods significantly outperform BeamformIt. The GEV beamformer endowed with the pretrained VAD achieves the best results. It should be

[4] Microphone 2 is not used in the case of the real-world recordings as, here, it is oriented away from the speaker.



(a) Percentage histogram of SINR improvement over full dataset.

(b) Percentage histogram of SINR improvement over the subset with small movements of the SOI.

Fig. 6 Histograms of SINR improvement achieved by the variants of CSV-AuxIVE in the experiment of Section 4.3

noted that the rates achieved by the BSE techniques are comparable to GEV even without a training stage on any CHiME-4 data.

In general, the block-wise methods achieve lower WER than their counterparts based on the static mixing model; the WER of BOGIVE$_\mathbf{w}$ is comparable with CSV-AuxIVE. A significant advantage of the latter method is the faster convergence and, consequently, much lower computational burden. The total duration of the 5920 files in the CHiME-4 dataset is 10 h and 5 min. The results presented for BOGIVE$_\mathbf{w}$ have been achieved after 100 iterations on each file, which translates into 10 hours and 30 minutes[5] of processing for the whole dataset. CSV-AuxIVE is able to converge in 7 iterations; the whole enhancement was finished in 1 h and 2 min.

An example of the enhancement yielded by the block-wise methods on one of the CHiME-4 recordings is shown in Fig. 7. Within this particular recording, in the interval 1.75–3 s, the target speaker was moved out of its initial position. The OverIVA algorithm focused on this initial direction only, resulting in vanishing voice during the movement interval. Consequently, the automatic transcription is erroneous. In contrast, CSV-AuxIVE is able to focus on both positions of the speaker and recovers the signal of interest correctly. The fact that there are few such recordings with significant speaker movement in the CHiME-4 datasets explains why the achieved improvements of WER by the block-wise methods are small.

5 Conclusions

The ability of the CSV-based BSE algorithms to extract moving acoustic sources has been corroborated by the experiments presented in this paper. The blind extraction is based on the estimation of a separating filter that passes signals from the entire area of the source presence. This way, the moving source can be extracted efficiently without tracking in an on-line fashion. The experiments show that these methods are particularly robust with respect to small source movements and effectively exploit overdetermined settings, that is, when there is a higher number of microphones than that of the sources.

We have proposed a new BSE algorithm of this kind, CSV-AuxIVE, which is based on the auxiliary function-based optimization. The algorithm was shown to be faster in convergence compared to its gradient-based counterpart. Furthermore, we have proposed the semi-supervised variant of CSV-AuxIVE utilizing pilot signals. The experiments confirm that this algorithm yields stable global convergence to the SOI.

Table 5 WERs [%] achieved in the CHiME-4 challenge

System	Development		Test	
	REAL	SIMU	REAL	SIMU
Unprocessed	9.83	8.86	19.90	10.79
BeamformIt	5.77	6.76	11.52	10.91
GEV (VAD)	**4.61**	**4.65**	**8.10**	**5.99**
OGIVE$_\mathbf{w}$	5.59	4.96	9.51	6.34
BOGIVE$_\mathbf{w}$	5.49	4.91	9.19	6.44
OverIVA	5.97	5.21	10.43	6.82
CSV-AuxIVE	5.65	4.83	9.88	6.46

[5] The computations run on a workstation using an Intel i7-2600K@3.4GHz processor with 16GB RAM.

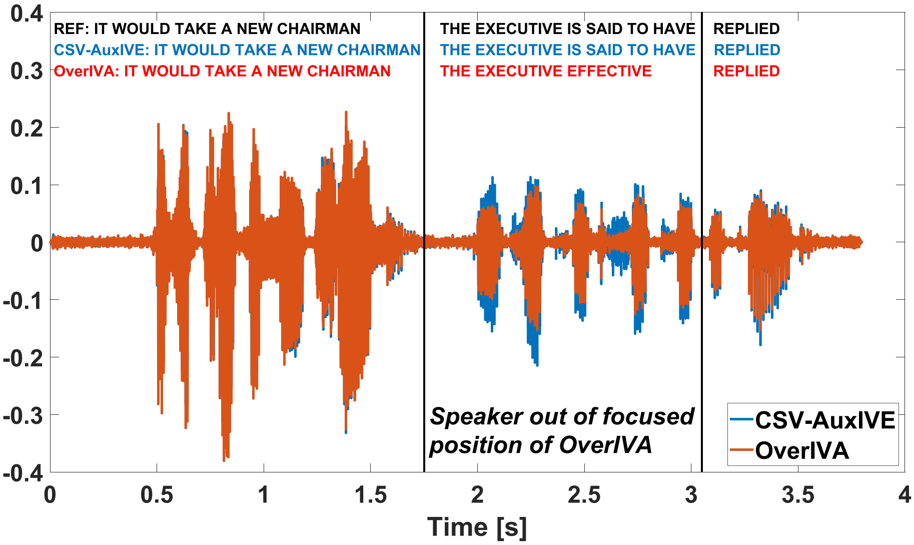


Fig. 7 Comparison of enhanced signals yielded from a recording of a moving speaker by OverIVA and CSV-AuxIVE

For the future, the proposed methods provide us with alternatives to the conventional approaches that adapt to the source movements through application of static mixing models on short time-intervals. Their other abilities, for example, the adaptability to high speed speaker movements and the robustness against a highly reverberant and noisy environment, pose an interesting topic for future research [35].

Abbreviations

BSS: Blind source separation; BSE: Blind source extraction; ICA: Independent component analysis; STFT: Short-time fourier transform; FDICA: Frequency-domain ICA; IVA: Independent vector analysis; ILRMA: Independent low rank matrix analysis; NMF: Nonnegative matrix factorization; ICE: Independent component extraction; IVE: Independent vector extraction; CSV: Constant separating vector; SOI: Signal of interest; ISR: Interference-to-signal ratio; OGC: Orthogonal constraint; AFO: Auxiliary function-based optimization; (D&S): Delay-and-sum; IR: Interference; SIR: Signal-to-interference ratio; SDR: Signal-to-distortion ratio; iSIR: improvement in signal-to-interference ratio; iSDR: improvement in signal-to-distortion ratio; nSIR: normalized signal-to-interference ratio; OOG: Out-of-grid position; FFT: Fast fourier transform; RIR: Room impulse response; iSINR: improvement in signal-to-interference-and-noise ratio; WER: Word error rate; GEV: Generalized eigenvalue beamformer; VAD: Voice-activity-detector; BAN: Blind analytic normalization; RTF: Relative transfer function

Authors' contributions

JJ designed the proposed method, evaluated the experiments and wrote the paper (except 1). ZK wrote Section 1 and provided paper correction. JM provided experiments concerning CHiME-4 dataset in Section 4.4. TK prepared data for experiments (4.2, 4.3) and provided final text correction. JČ prepared data for experiments described in 4.2 and 4.3 and edited the tables and figures. All the authors read and approved the final manuscript.

Funding

This work was supported by The Czech Science Foundation through Projects No. 17-00902S and No. 20-17720S, by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-19-1-2105, and by the Student Grant Competition of the Technical University of Liberec under the project No. SGS-2019-3060.

Availability of data and materials

Dataset and results from Section 4.2 are available at: https://asap.ite.tul.cz/downloads/ice/blind-extraction-of-a-moving-speaker/
MIRaGe database with it's additional support software (used for Section 4.3) is available at: https://asap.ite.tul.cz/downloads/mirage/
CHiME-4 dataset from Section 4.4 is publicaly available at: http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 29 April 2021 Accepted: 14 November 2021
Published online: 04 January 2022

References

1. S. Makino, T.-W. Lee, H. Sawada (eds.), *Blind speech separation*, vol. 615 (Springer, Dordrecht, 2007)
2. P. Comon, C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications. Independent Component Analysis and Applications Series*. (Elsevier Science, Amsterdam, 2010)
3. E. Vincent, T. Virtanen, S. Gannot, *Audio source separation and speech enhancement*, 1st edn. (Wiley Publishing, Chichester, 2018)
4. A. Hyvärinen, J. Karhunen, E. Oja, *Independent component analysis*. (John Wiley & Sons, Chichester, 2001)
5. P. Comon, Independent component analysis, a new concept?. Sig. Process. **36**, 287–314 (1994)
6. P. Smaragdis, Blind separation of convolved mixtures in the frequency domain. Neurocomputing. **22**, 21–34 (1998)
7. H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Process. **12**(5), 530–538 (2004)
8. T. Kim, I. Lee, T. Lee, in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*. Independent vector analysis: definition and algorithms (IEEE, Piscataway, 2006), pp. 1393–1396
9. T. Kim, H. T. Attias, S.-Y. Lee, T.-W. Lee, in *IEEE Transactions on Audio, Speech, and Language Processing, vol. 15*. Blind source separation exploiting higher-order frequency dependencies (IEEE Press, 2007), pp. 70–79
10. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, H. Saruwatari, Determined blind source separation unifying independent vector analysis and

nonnegative matrix factorization. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(9), 1626–1641 (2016)

11. D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, K. Kondo, Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation. EURASIP J. Adv. Sig. Process. **2018**(1), 28 (2018)
12. N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, N. Ono, Independent deeply learned matrix analysis for determined audio source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(10), 1601–1615 (2019). https://doi.org/10.1109/TASLP.2019.2925450
13. Z. Koldovský, P. Tichavský, Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence. IEEE Trans. Sig. Process. **67**(4), 1050–1064 (2019)
14. R. Scheibler, N. Ono, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Fast independent vector extraction by iterative SINR maximization (IEEE, Piscataway, 2020), pp. 601–605
15. R. Ikeshita, T. Nakatani, Independent Vector Extraction for Joint Blind Source Separation and Dereverberation (2021). 2102.04696
16. R. Mukai, H. Sawada, S. Araki, S. Makino, in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*. Robust real-time blind source separation for moving speakers in a room, vol. 5, (2003), p. 469. https://doi.org/10.1109/ICASSP.2003.1200008
17. T. Taniguchi, N. Ono, A. Kawamura, S. Sagayama, in *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. An auxiliary-function approach to online independent vector analysis for real-time blind source separation (IEEE, Piscataway, 2014), pp. 107–111
18. A. H. Khan, M. Taseska, E. A. P. Habets, in *A Geometrically Constrained Independent Vector Analysis Algorithm for Online Source Extraction*, ed. by E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský (Springer, Cham, 2015), pp. 396–403
19. S.-H. Hsu, T. R. Mullen, T.-P. Jung, G. Cauwenberghs, Real-time adaptive eeg source separation using online recursive independent component analysis. IEEE Trans. Neural Syst. Rehabil. Eng. **24**(3), 309–319 (2016)
20. N. Ono, in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Stable and fast update rules for independent vector analysis based on auxiliary function technique (IEEE, Piscataway, 2011), pp. 189–192
21. T. Nakashima, R. Scheibler, Y. Wakabayashi, N. Ono, in *2020 28th European Signal Processing Conference (EUSIPCO)*. Faster independent low-rank matrix analysis with pairwise updates of demixing vectors, (2021), pp. 301–305. https://doi.org/10.23919/Eusipco47968.2020.9287508
22. R. Scheibler, N. Ono, in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Independent vector analysis with more microphones than sources (IEEE, Piscataway, 2019), pp. 185–189
23. Z. Koldovský, J. Málek, J. Janský, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*. Extraction of independent vector component from underdetermined mixtures through block-wise determined modeling (IEEE, Piscataway, 2019), pp. 7903–7907
24. V. Kautský, Z. Koldovský, P. Tichavský, V. Zarzoso, Cramér-Rao bounds for complex-valued independent component extraction: Determined and piecewise determined mixing models. IEEE Trans. Sig. Process. **68**, 5230–5243 (2020)
25. Z. Koldovský, P. Tichavský, V. Kautský, in *Proceedings of European Signal Processing Conference*. Orthogonally constrained independent component extraction: Blind MPDR beamforming (IEEE, Piscataway, 2017), pp. 1195–1199
26. K. Kreutz-Delgado, The complex gradient operator and the cr-calculus. arXiv (2009). 0906.4835
27. Z. Koldovský, F. Nesta, Performance analysis of source image estimators in blind source separation. IEEE Trans. Sig. Process. **65**(16), 4166–4176 (2017)
28. L. C. Parra, C. V. Alvino, Geometric source separation: merging convolutive source separation with geometric beamforming. IEEE Trans. Speech Audio Process. **10**(6), 352–362 (2002)
29. S. Bhinge, R. Mowakeaa, V. D. Calhoun, T. Adalı, Extraction of time-varying spatiotemporal networks using parameter-tuned constrained IVA. IEEE Trans. Med. Imaging. **38**(7), 1715–1725 (2019)
30. A. Brendel, T. Haubner, W. Kellermann, A unified probabilistic view on spatially informed source separation and extraction based on independent vector analysis. IEEE Trans. Sig. Process. **68**, 3545–3558 (2020)
31. F. Nesta, Z. Koldovský, in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*. Supervised independent vector analysis through pilot dependent components (IEEE, Piscataway, 2017), pp. 536–540
32. F. Nesta, S. Mosayyebpour, Z. Koldovský, K. Paleček, in *Proceedings of European Signal Processing Conference*. Audio/video supervised independent vector analysis through multimodal pilot dependent components (IEEE, Piscataway, 2017), pp. 1190–1194
33. J. Čmejla, T. Kounovský, J. Málek, Z. Koldovský, in *Latent Variable Analysis and Signal Separation*, ed. by Y. Deville, S. Gannot, R. Mason, M. D. Plumbley, and D. Ward. Independent vector analysis exploiting pre-learned banks of relative transfer functions for assumed target's positions (Springer, Cham, 2018), pp. 270–279
34. J. Janský, J. Málek, J. Čmejla, T. Kounovský, Z. Koldovský, J. Žd'ánský, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Adaptive blind audio source extraction supervised by dominant speaker identification using x-vectors (IEEE, Piscataway, 2020), pp. 676–680
35. J. Malek, J. Jansky, T. Kounovsky, Z. Koldovsky, J. Zdansky, in *Accepted for ICASSP2021*. Blind extraction of moving audio source in a challenging environment supported by speaker identification via X-vectors (IEEE, Piscataway, 2021)
36. J. B. Allen, D. A. Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
37. J. S. Garofolo, et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. (Linguistic Data Consortium, Philadelphia, 1993)
38. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
39. J. Čmejla, T. Kounovský, S. Gannot, Z. Koldovský, P. Tandeitnik, in *Proceedings of European Signal Processing Conference*. Mirage: Multichannel database of room impulse responses measured on high-resolution cube-shaped grid in multiple acoustic conditions (IEEE, Piscataway, 2020), pp. 56–60
40. E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, R. Marxer, An analysis of environment, microphone and data simulation mismatches in robust speech recognition. Comput. Speech Lang. **46**, 535–557 (2017). https://doi.org/10.1016/j.csl.2016.11.005
41. Z. Koldovský, J. Málek, P. Tichavský, F. Nesta, Semi-blind noise extraction using partially known position of the target source. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2029–2041 (2013)
42. J. Málek, Z. Koldovský, M. Boháč, Block-online multi-channel speech enhancement using dnn-supported relative transfer function estimates. IET Sig. Process. **14**, 124–133 (2020)
43. X. Anguera, C. Wooters, J. Hernando, Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Lang. Process. **15**(7), 2011–2022 (2007)
44. E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, R. Marxer, The 4th CHiME Speech Separation and Recognition Challenge. http://spandh.dcs.shef.ac.uk/chime_challenge/chime2016/. Accessed 02 Dec 2019
45. J. Heymann, L. Drude, R. Haeb-Umbach, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Neural network based spectral mask estimation for acoustic beamforming (IEEE, Piscataway, 2016), pp. 196–200
46. J. Heymann, L. Drude, R. Haeb-Umbach, in *Proc. of the 4th Intl. Workshop on Speech Processing in Everyday Environments, CHiME-4*. Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition, (2016)
47. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Sig. Process. **49**(8), 1614–1626 (2001). https://doi.org/10.1109/78.934132

Publisher's Note