Multichannel Blind Sound Source Separation using Spatial Covariance Model with Level and Time Differences and Non-Negative Matrix Factorization

J.J. Carabias-Orti, J. Nikunen, T. Virtanen, *Senior Member, IEEE* and P. Vera-Candeas

***Abstract*—This paper presents an algorithm for multichannel sound source separation using explicit modeling of level and time differences in source spatial covariance matrices (SCM). We propose a novel SCM model in which the spatial properties are modeled by the weighted sum of direction of arrival (DOA) kernels. DOA kernels are obtained as the combination of phase and level difference covariance matrices representing both time and level differences between microphones for a grid of predefined source directions. The proposed SCM model is combined with the NMF model for the magnitude spectrograms. Opposite to other SCM models in the literature, in this work, source localization is implicitly defined in the model and estimated during the signal factorization. Therefore, no localization preprocessing is required. Parameters are estimated using complex-valued non-negative matrix factorization (CNMF) with both Euclidean distance and Itakura Saito divergence. Separation performance of the proposed system is evaluated using the two-channel SiSEC development dataset and four channels signals recorded in a regular room with moderate reverberation. Finally, a comparison to other state-of-the-art methods is performed, showing better achieved separation performance in terms of SIR and perceptual measures.**

***Index Terms*—multichannel source separation, spatial covariance model, interaural time difference, interaural level difference, non-negative matrix factorization, direction of arrival estimation.**

I. Introduction

In the context of audio signal processing, sound source separation aims at recovering each source signal from a set of audio mixtures of the original sources, such as those obtained by a microphone array, a binaural recording or an audio CD. Source separation has been widely applied for several audio processing tasks, including music remixing [1], automatic karaoke [2], instrument-wise equalization [3] or music information retrieval systems [4].

In this work, we propose a method for blind source separation (BSS). The term blind is used to emphasize that very little information about the sources or the mixing process is known. A common approach to this type of problem is based on independent component analysis (ICA), in which the underlying source signals are constrained to be statistically independent and non-Gaussian [5]. Unfortunately, ICA-based methods are subject to the well-known scale and source permutation ambiguity (i.e. the energies and order of the sources cannot be determined). Several methods in the literature have used time difference of arrival (TDOA) to interpret the ICA mixing parameters [6], [7], [8]. These methods assume the multichannel audio file corresponds to the recorded channels from a microphone array with known configuration. Other methods use the information from the TDOAs between microphones to create time-frequency masks to cluster all the time-frequency points with similar spatial properties [9], [10].

Beamforming techniques can be also applied for BSS obtaining satisfactory perceptual results [11], although the isolation of the target sources is limited. To improve the separation results, several works propose to use postfiltering techniques (see [12] for a review).

More recent methods are based on non-negative matrix factorization (NMF). However, while most music recordings are available in multichannel format (commonly, stereo[1]), standard NMF is only suited to single-channel data. Extensions to multichannel data have been considered, either by stacking up the spectrograms of all the channels into a single matrix [13] or by using nonnegative tensor factorization (NTF) under a parallel factor analysis (PARAFAC) structure, where the channel spectrograms form the slices of a 3-valence tensor [14], [15], [16]. In these approaches, the multichannel spectrogram is modeled by linear combination of the individual source magnitude (or power) spectra, which approximates the instantaneous mixing in time domain only if the sources have identical phase spectra and under anechoic conditions. Another interesting extension is complex non-negative matrix factorization (CNMF) [17], which employs a factorization-type model for the magnitude of a time-frequency representation similar to NMF but additionally estimates a phase matrix for each source. The phase information can be used to improve the separation quality of overlapping partials, especially when phase cancellation occurs [18]. However, due to the size and number of elements in the phase matrices, the number of free parameters in CNMF is considerably higher compared to NMF, which in practice can lead to poor local minima during

Manuscript received XX XX, XXXX; revised XX XX, XXXX; accepted XX XX, XXXX. Date of publication XX XX, XXXX; date of current version XX XX, XXXX. This work has been funded by the Academy of Finland project number 290190. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Huseyin Hacihabiboglu. (Corresponding author: Julio Carabias.)

J.J. Carabias-Orti, J. Nikunen and T. Virtanen are with the Department of Signal Processing, Tampere University of Technology, Tampere 33720, Finland (e-mail: carabiasjulio@gmail.com; joonas.nikunen@tut.fi; tuomas.virtanen@tut.fi).

P. Vera-Candeas is with the Department of Telecommunication Engineering, University of Jaen, Jaen, Spain. (e-mail: pvera@ujaen.es).

[1]Stereo signals relate to the loudspeaker channels that are typically created through artistic mixing of recorded sources.

model fitting.

Another way of modeling the spatial properties is based on a spatial covariance matrix (SCM) as signal representation [19], [20], [21], [22]. For each time-frequency point in the STFT, the SCM represents the mixing of the sources by magnitude correlations and phase differences between channels. Opposite to the complex NMF model in [17], it is not dependent on the absolute phase of the source signal.

The CNMF algorithms in [19], [20] estimate unconstrained SCM mixing filters together with NMF magnitude (or power) model to identify and separate repetitive frequency patterns corresponding to a single spatial location. Another strategy of estimating the covariance matrices is Gaussian modeling. For the task of source separation, studies [25] and [26] proposed to estimate the SCMs using iterative EM algorithm together with a multichannel Wiener filter to extract sources from the mixture. The mixing model is assumed to be frequency dependent and no cross-frequency information is utilized in SCM parameter estimation. Recently, the NMF spectrogram model has been replaced by deep learning strategies, and use of deep neural networks (DNNs) for modeling the source spectrogram in combination with Gaussian SCM model was proposed in [27], [28] where it was reported to outperform NMF-based models.

When dealing with spectrally similar sources (e.g. several speech signals), without further constraints, NMF-based approaches can lead to the situation where a single NMF component together with the corresponding SCM mixing filter represent multiple sources together at different spatial locations. To enforce SCMs at different frequencies to corresponds to the same location, Nikunen and Virtanen [21], [22] proposed a SCM model based on DOA kernels that represent the phase difference between microphones caused by a single spatial location and its analytic TDOA for a given array geometry. Thus, the SCMs of each source were modeled as a weighted combination of the DOA kernels scanning all possible directions of arrivals for sources. The DOA-based SCM model only requires estimation of frequency-independent directional weights. As a result, the effect of the spatial aliasing is mitigated since the model accounts for phase difference evidence across frequency by single frequency-independent time delays of individual DOA kernels.

Several source localization and DOA estimation methods [29], [30] also assume that the soundfield can be represented as a spatially sparse distribution of sound sources over an overcomplete linear equation of the observations. However, in those methods no assumptions are imposed on the structure of the source signals in the time-frequency domain. Alternatively, NMF approximates the time-frequency spectrogram of the source signal as the product of two nonnegative low-rank matrices, allowing further constraints on the source signal model. For instance, a recent approach in [31] combines supervised CNMF and sparse sound field decomposition, obtaining superior results for DOA estimation than other methods using sparse representations.

The SCM methods in [21] and [22] have three limitations. First, DOA kernels account for time difference between array channels (using phase differences), while level (or intensity) differences between channels are estimated during the factorization, independently for each frequency. Consequently, a large number of parameters has to be tuned and thus, without any prior information, these methods are prone to converge to local minima. Second, the source reconstruction requires a post-processing clustering stage to group the NMF components together with their associate SCM mixing filters to sources, or as in [22], prior information about the spatial cues before associating components to sources. Third, both methods used CNMF with Euclidean distance (EUC) whereas other cost functions such as Itakura Saito (IS) divergence are better suited for audio modeling [32].

In this paper, we propose a novel SCM-based model and a constrained CNMF algorithm for BSS that enables to estimate both source localization and separation jointly during the factorization, without the need of any prior information about the source location nor post-processing stage. The main contributions of this paper are summarized as follows. 1) A SCM kernel based model where the mixing filter is decomposed into two direction dependent SCMs to represent both time and level differences between array channels. The level differences are represented using a panning-inspired frequency-independent covariance matrix. On the contrary, time delays are modeled using a frequency-dependent phase difference covariance matrix. The main benefit of explicitly modeling the panning-inspired level-difference covariance matrix is the reduction of the number of free parameters avoiding non-coherent level differences between channels across frequency. 2) Two novel group sparsity constraints for source localization that enforce non-overlapping DOAs between sources and a single DOA for each source. 3) Algorithms for minimizing the IS divergence between the SCM kernel based CNMF model and the observations.

In order to evaluate our approach, we have used the two channels recordings from the SiSEC'08 [33] development dataset. Moreover, to test our system performance with more sensors, we have created a multichannel dataset using impulse responses (IR) captured with two microphone arrays in a reverberant room and convolving the anechoic signals from SiSEC'08 development dataset with the IRs to create mixtures of two and three sources. Both microphone arrays consist of four microphones with 5 cm and 54 cm distance between sensors, respectively. A comparison to other multichannel methods is performed demonstrating the reliability and robustness of our proposal independently of the recording conditions.

The rest of the article is organized as follows. The signal mixing representation using spatial covariance matrices (SCM) is presented Section II. Section III describes the proposed SCM model using both time and level differences between channels. Formulation of the proposed SCM model into the CNMF framework and update rules for the model parameter optimization are presented in Section IV. Section V describes the source reconstruction method. Experimental setup and evaluation of the proposed method is performed in Section VI. Finally, conclusions are presented in section VII together with a discussion about future work.

II. Signal Model using Spatial Covariance Matrices

In this section we define the problem of the sound source separation with spatial audio captures and present the spatial processing background for the proposed SCM model and CNMF algorithm it is used with. The section consists of the source mixing model in Section II-A, definition of the signal representation and the spatial covariance matrices in Section II-B and interpretation of the convolutive mixing model in the spatial covariance domain in Section II-C.

A. Source Mixing Model

The mixing model for a multichannel recording can be represented as the convolution of each sound source $y_s(n)$ in the mixture with its corresponding spatial room impulse response $h_{ms}(n)$ as

$$x_m(n) = \sum_{s=1}^{S} \sum_{\tau} h_{ms}(\tau) y_s(n-\tau), \tag{1}$$

where mixture $x_m(n)$ consist of $s = 1...S$ sources captured by $m = 1...M$ microphones, n being the time sample index.

In the short-time Fourier transform (STFT) domain, the mixing model in Eq. (1) can be approximated by an instantaneous mixing at each time-frequency (f, t) point as

$$\tilde{\mathbf{x}}_{ft} \approx \sum_{s=1}^{S} \tilde{\mathbf{h}}_{fs} \tilde{y}_{fts}, \tag{2}$$

where $\tilde{\mathbf{x}}_{ft} = [\tilde{x}_{ft1}, ..., \tilde{x}_{ftM}]^T \in \mathbb{C}^M$ is the STFT of the multichannel mixture $x_m(n)$ at frequency bin f and time frame t. $\tilde{y}_{fts}$ is the complex-valued STFT of the monaural source $s \in [1...S]$. The mixing filter is defined in frequency domain as $\tilde{\mathbf{h}}_{fs} = [\tilde{h}_{fs1}, ..., \tilde{h}_{fsM}]^T \in \mathbb{C}^M$. Despite the effective length of the spatial/room impulse response $h_{ms}(n)$ can be several hundreds milliseconds, in practice, a shorter STFT analysis window of tens of milliseconds is enough to capture the direct sound and the main reverberant part [21].

B. Signal Representation

In this work we use the spatial covariance matrix (SCM) signal representation used in [20], [21]. Rather than absolute phase values, a SCM represents the phase difference between every pair of microphones in the multichannel mixture.

For each frequency bin f and time frame t, the magnitude square-rooted matrix $\tilde{\mathbf{x}}_{ft}$ of the captured signal at each sensor $\tilde{\mathbf{x}}_{ft} = [\tilde{x}_{ft1}, ..., \tilde{x}_{ftM}]^T$ is given by

$$\hat{\mathbf{x}}_{ft} = [|\tilde{x}_{ft1}|^{1/2} sgn(\tilde{x}_{ft1}), ..., |\tilde{x}_{ftM}|^{1/2} sgn(\tilde{x}_{ftM})]^T, \tag{3}$$

where $sgn(z) = z/|z|$ is the signum function for complex numbers. Then, the SCM $\mathbf{X}_{ft}$ for a single time-frequency point (f, t) is defined from the array captured signal $\tilde{\mathbf{x}}_{ft}$ (see Eq. (2)) as the outer product

$$\mathbf{X}_{ft} = \hat{\mathbf{x}}_{ft} \hat{\mathbf{x}}_{ft}^H = \begin{bmatrix} |\tilde{x}_{ft1}| & \cdots & \tilde{x}_{ft1} \tilde{x}_{ftM}^* \\ \vdots & \ddots & \vdots \\ \tilde{x}_{ftM} \tilde{x}_{ft1}^* & \cdots & |\tilde{x}_{ftM}| \end{bmatrix}, \tag{4}$$

where ^{H} stands for Hermitian transpose. For each time-frequency point (f, t), the diagonal of $\mathbf{X}_{ft}$ represents the non-negative real-valued magnitude of the observation at each microphone $[|\tilde{x}_{ft1}|, ..., |\tilde{x}_{ftM}|]^T$. On the contrary, the off-diagonal values of $[\mathbf{X}_{ft}]_{pm}$ with $p \neq m$ represent the magnitude correlation and phase difference $|\tilde{x}_{ftp} \tilde{x}_{ftm}|^{1/2} sgn(\tilde{x}_{ftp} \tilde{x}_{ftm}^*)$ between microphone pair (p, m).

C. Spatial Covariance Source Mixing Model

The source mixing model in Eq. (2) can be approximated in terms of the SCM representation as

$$\mathbf{X}_{ft} \approx \hat{\mathbf{X}}_{ft} = \sum_{s=1}^{S} \mathbf{H}_{fs} \overline{y}_{fts}, \tag{5}$$

where $\mathbf{H}_{fs} \in \mathbb{C}^{M \times M}$ is the SCM representation of the spatial frequency response $\tilde{\mathbf{h}}_{fs}$ and $\overline{y}_{fts} = |\tilde{y}_{fts}|$ is the magnitude spectrum for each source $s \in [1, ..., S]$. As explained in [21], we can approximate the SCM model in Eq. (5) to be purely additive since the sources are approximately uncorrelated and sparse (i.e. only a single source is active at each time frequency (f, t) point). Note that, despite the sparsity assumption is often used in the existing research it does not always hold in practice, particularly in reverberant environments.

III. Proposed Level/Time Spatial Covariance Matrix Model for BSS

In this work, we propose an extension of the SCM-based CNMF signal model proposed in [22]. In particular, we propose to decompose the DOA kernels into a combination of phase and level covariance matrices to account for both time and level difference between array channels.

Moreover, we propose a novel strategy to perform source separation. As in [20], [22], we propose to relate NMF components to sources to avoid using a post-processing clustering stage but, opposite to [22], no prior information about the source directions is required here. Instead, we impose two constraints to the spatial weights of the sources to avoid overlapping source directions and to enforce a single direction per source.

The block diagram of the proposed method is displayed in Figure 1. First the SCM representation in Eq. (4) is computed from the STFT of the multichannel mixture. Second, the model parameters are estimated using a SCM kernel-based CNMF algorithm in two steps: 1) Initialization of the localization and source spectrogram parameters using two novel single/non-overlapping direction constraints, 2) Estimation of the magnitude spectrogram and the panning mixing parameters. Finally, a generalized Wiener filtering strategy is used to obtain the source reconstruction.

A. Proposed Multichannel Signal Model

The SCM mixing filter $\mathbf{H}_{fs}$ in Eq. (5) accounts for amplitude and phase differences between channels, however it does not have any explicit relation to spatial locations.

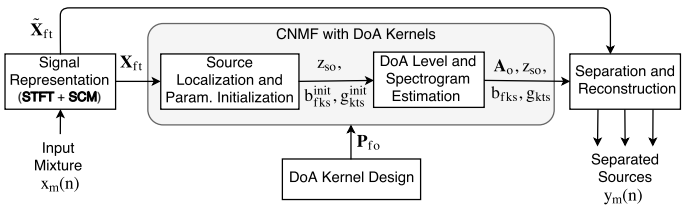

Fig. 1. Block diagram of the proposed SCM-based BSS system

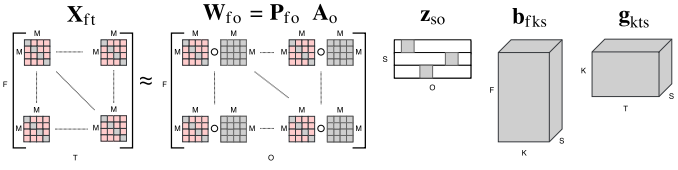

Fig. 2. Proposed signal SCM model parameters. Complex values are displayed in red, positive real values in gray and zero values in white.

Beamforming-inspired SCM methods in [21] and [22] model the SCM mixing filter $\mathbf{H}_{fs}$ as a linear combination of the DOA kernels $\mathbf{W}_{fo}$ multiplied by the spatial weights matrix $z_{ko} \in \mathbb{R}^{\geq 0}$ which relates NMF components k with spatial directions o. However, due to the amount of free parameters, these methods are prone to localization errors when no prior information is given.

In this work, we propose a SCM-based model that enables to estimate the spatial location/position and the spectrogram of the sources jointly during the factorization, without the need of prior information nor any post-processing stage. The proposed signal model for SCM observation is presented in Eq. (6) as

$$\mathbf{X}_{ft} \approx \hat{\mathbf{X}}_{ft} = \sum_{s=1}^{S} \underbrace{\sum_{o=1}^{O} \overbrace{(\mathbf{P}_{fo} \circ \mathbf{A}_o)}^{\mathbf{W}_{fo}} z_{so}}_{\mathbf{H}_{fs}} \underbrace{\sum_{k=1}^{K} b_{fks} g_{kts}}_{\overline{y}_{fts}}, \quad (6)$$

where $\circ$ stands for the element-wise multiplication (i.e. Hadamard product). The model in Eq. (6) is illustrated in Figure 2.

To reduce the number of free parameters and enforce the coherence of the relative amplitude (i.e. level differences) between channels across frequency, we propose to decompose the DOA kernels $\mathbf{W}_{fo}$ into two covariance matrices, the phase differences covariance matrix (PDCM) $\mathbf{P}_{fo}$ and the level differences covariance matrix (LDCM) $\mathbf{A}_o$. Previous approaches using beamforming-inspired SCM mixing models in [21] and [22] keep the phase of $\mathbf{W}_{fo}$ fixed and update only its magnitudes (i.e. the computed phase update is discarded). Alternatively, in our approach, the relative amplitudes between microphones are estimated using the LDCM while the PDCM is kept fixed during the factorization.

First, the frequency dependent PDCM represents the phase difference between microphones. In fact, $\mathbf{P}_{fo} \in \mathbb{C}^{M \times M}$ is computed a priori for every spatial position as

$$[\mathbf{P}_{fo}]_{pm} = \exp(\mathrm{j}\theta_{\mathrm{p,m}}(\mathrm{f,o})), \quad (7)$$

where $\theta_{p,m}(f,o) = 2\pi f_i \tau_{pm}(\mathbf{r}_o)$ represents the phase difference computed from the TDOA between sensors p and m for the frequency in Hz at bin f and the spatial position o. Note that a fixed number of look directions $o = 1, ..., O$ are used to cover the ranges of $-90^o \leq \theta \leq 90^o$ and $0^o \leq \phi \leq 360^o$ in elevation and azimuth, respectively. Each spatial position $\mathbf{r}_o$ can be translated to a TDOA (in seconds) for a pair of microphones (p, m) using the following expression:

$$\tau_{pm}(\mathbf{r}_o) = \frac{\|\mathbf{r}_o - \mathbf{p}\|_2 - \|\mathbf{r}_o - \mathbf{m}\|_2}{c}, \quad (8)$$

where $\|\cdot\|_2$ denotes the L2-norm, $\mathbf{r}_o$, $\mathbf{m}$ and $\mathbf{p}$ are the source spatial location/position corresponding to direction o, and the microphone m and p locations using the Cartesian coordinate system, respectively, and c is the speed of sound.

Second, assuming anechoic conditions and that the level of the captured signal varies as a function of the direction of arrival and microphone directivity pattern, the frequency independent LDCM $\mathbf{A}_o \in \mathbb{R}^{M \times M}$ represents the relative level factor (i.e. panning) between sensors. Actually, the LDCM is initialized as $[\mathbf{A}_o]_{nm} = 1/M$, where M is the number of channels in the array but, opposite to the PDCM, the LDCM is a free parameter to be estimated during the factorization.

Therefore, for each frequency and direction pair (f, o), the DOA kernel $\mathbf{W}_{fo} \in \mathbb{C}^{M \times M}$ is obtained as the element-wise multiplication of the unit-amplitude complex-valued PDCM $\mathbf{P}_{fo}$ and the zero-phase real-valued LDCM $\mathbf{A}_o$ as depicted in Eq. (9) (see top of next page).

Finally, for each source s the magnitude (or power) time-frequency spectrogram $\overline{y}_{fts}$ is obtained as a linear combination of NMF basis functions b_{fks} and their corresponding time-varying gains g_{kts} (see Figure 2). In other SCM-based methods in the literature [20], [21] the NMF components are associated to sources using a post-processing clustering of their spatial weights. To avoid this clustering, in [22] the authors proposed a soft-decision parameter to relate NMF components to sources but without a prior information the method is prone to converge to a local minima. Here, the spatial weights matrix $z_{so} \in \mathbb{R}^{\geq 0}$ is explicitly defined to relate sources s with spatial directions o without the need of any clustering nor intermediate (i.e. soft-decision) parameter (see Figure 2).

The proposed SCM model in Eq. (6) is less prone to localization errors than the beamforming-inspired SCM model in [21] and [22] for three reasons: 1) The number of free parameters is lower which contributes to the robustness of the method, 2) The LDCM enforces the coherence of the relative amplitudes between microphones across frequency, 3) A NMF component is associated to a single source which favors the independence and orthogonality of the learned NMF components.

B. Source Localization using Penalty Terms

Without further constraints, the estimation of the spatial position for each source using the proposed model in Eq. (6) is prone to ambiguity, i.e. the spatial weights may not be sparse. To overcome this problem, in [22] the authors proposed a hard prior initialization of the spatial weights z_{so} using

$$\mathbf{W}_{fo} = \underbrace{\begin{bmatrix} 1 & \mathrm{e}^{j\theta_{12}(f,o)} & \cdots & \mathrm{e}^{j\theta_{1m}(f,o)} \\ \mathrm{e}^{j\theta_{21}(f,o)} & 1 & \cdots & \mathrm{e}^{j\theta_{2M}(f,o)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{e}^{j\theta_{M1}(f,o)} & \mathrm{e}^{j\theta_{M2}(f,o)} & \cdots & 1 \end{bmatrix}}_{\mathbf{P}_{fo}} \circ \underbrace{\begin{bmatrix} a_1(o) & \sqrt{a_1(o)a_2(o)} & \cdots & \sqrt{a_1(o)a_M(o)} \\ \sqrt{a_2(o)a_1(o)} & a_2(o) & \cdots & \sqrt{a_2(o)a_M(o)} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{a_M(o)a_1(o)} & \sqrt{a_M(o)a_2(o)} & \cdots & a_M(o) \end{bmatrix}}_{\mathbf{A}_o} \tag{9}$$

the information from the steered response power (SRP) [34] algorithm for source localization.

However, in this work we aim to estimate the spatial position for each source without any prior information nor post-processing stage. To this end, we propose to constrain the proposed model in Eq. (6) to satisfy two conditions: 1) assuming moderate reverberation, a source arrives from a single direction (i.e. the spatial weights should be sparse), 2) a single spatial position cannot be assigned to multiple sources.

To enforce the spatial weights to be sparse and to avoid overlapping spatial positions between sources, two group sparsity/cross-correlation-based regularization terms $\varphi_1(\mathbf{Z})$ and $\varphi_2(\mathbf{Z})$ are presented. $\mathbf{Z}$ stands for the matrix notation of the spatial weights z_{so}.

First, $\varphi_1(\mathbf{Z})$ introduces a penalty for cross-correlation nonzero values between sources $(\mathbf{Z}\mathbf{Z}^T) \in \mathbb{R}^{S\times S}$ in the spatial weight matrix. In other words, $\varphi_1(\mathbf{Z})$ prevents multiple sources at the same direction and is defined as

$$\varphi_1(\mathbf{Z}) = \left\|\mathbf{C} \circ (\mathbf{Z}\mathbf{Z}^T)\right\|_1, \tag{10}$$

where $\|\cdot\|_1$ denotes the L1-norm and $\mathbf{C} \in \mathbb{R}^{S\times S}$ is a weighting matrix that selects which cross-terms to penalize and by how much. In this work, the weights are set such that the non-cross terms (elements on the diagonal) are not penalized, i.e. $C_{ii} = 0$, and the off-diagonal terms C_{ij} where $i \neq j$ are set to one. Note that a similar approach was used in [35], [36] to control the activation of musical notes in a music transcription task.

Second, under the assumption that a source arrives from a single direction, we enforce the spatial weights to be sparse by means of a single direction penalty $\varphi_2(\mathbf{Z})$ that penalize off-diagonal values in $(\mathbf{Z}^T\mathbf{Z}) \in \mathbb{R}^{O\times O}$ and thus, restricts the spatial weights to have only one predominant direction per source. The single direction penalty term is defined as

$$\varphi_2(\mathbf{Z}) = \left\|\mathbf{D} \circ (\mathbf{Z}^T\mathbf{Z})\right\|_1, \tag{11}$$

where $\mathbf{D} \in \mathbb{R}^{O\times O}$ is a Toeplitz matrix that, in this work, has been experimentally defined as $D_{xy} = \min(1, \log(1 + \frac{1}{5}|x - y|))$ so that, maximal penalty is applied to the weights corresponding to directions above ten degrees from the estimated source location.

IV. Complex-valued Non-Negative Matrix Factorization

In this work, we used CNMF to estimate the parameters of the proposed SCM model in Eq. (6). In previous works on multichannel NMF, expectation-maximization (EM) algorithms have been derived for both Euclidean (EUC) [19], [21], [22] and Itakura Saito (IS) [20] cost functions. Note that IS is better suited for audio modeling in comparison to EUC [32]. Alternatively, in this work, we present a similar approach to [20] to obtain the multiplicative updates via auxiliary functions for the case of the Euclidean (EUC) distance and the Itakura Saito (IS) divergence. In fact, as demonstrated in [20], these updates provide faster convergence than the EM algorithms.

A. Formulation for Euclidean Distance

As in [20], we aim to minimize the squared Frobenius norm between the observed $\mathbf{X}_{ft}$ and the estimated $\hat{\mathbf{X}}_{ft}$ SCM signal representations as follows:

$$D_{EUC}(\mathbf{X}_{ft}, \hat{\mathbf{X}}_{ft}) = \left\|\mathbf{X}_{ft} - \hat{\mathbf{X}}_{ft}\right\|_F^2, \tag{12}$$

In fact, the Euclidean distance in Eq. (12) together with the proposed DOA-based SCM model in Figure 2 can be expressed as:

$$f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G}) = -\sum_{f,t,s,o} z_{so}b_{fks}g_{kts}\,\mathrm{tr}(\mathbf{X}_{ft}(\mathbf{P}_{fo}\circ\mathbf{A}_o)^H) - \sum_{f,t,s,o} z_{so}b_{fks}g_{kts}\,\mathrm{tr}((\mathbf{P}_{fo}\circ\mathbf{A}_o)\,\mathbf{X}_{ft}^H) + \sum_{f,t}\mathrm{tr}\left(\hat{\mathbf{X}}_{ft}\hat{\mathbf{X}}_{ft}^H\right), \tag{13}$$

where constant terms are omitted. $\mathrm{tr}(\mathbf{X}) = \sum_{m=1}^{M} x_{mm}$ is the trace of a square matrix $\mathbf{X}$. To minimize the function in Eq. (13) we follow the optimization scheme of majorization in [20] using an auxiliary function f^+ as explained in Appendix A. Then, as in [20], the derivation of the algorithm updates is achieved via partial derivation of function f^+ w.r.t each model parameter and setting these derivatives to zero.

The update rules for the time-frequency spectrogram parameters in Eq. (6) are

$$b_{fks} \leftarrow b_{fks}\frac{\sum_{o,t} z_{so}g_{kts}\mathrm{tr}(\mathbf{X}_{ft}\mathbf{H}_{fs})}{\sum_{o,t} z_{so}g_{kts}\mathrm{tr}(\hat{\mathbf{X}}_{ft}\mathbf{H}_{fs})} \tag{14}$$

$$g_{kts} \leftarrow g_{kts}\frac{\sum_{o,f} z_{so}b_{fks}\mathrm{tr}(\mathbf{X}_{ft}\mathbf{H}_{fs})}{\sum_{o,f} z_{so}b_{fks}\mathrm{tr}(\hat{\mathbf{X}}_{ft}\mathbf{H}_{fs})} \tag{15}$$

Then, as explained in Section III-A, the SCM mixing filter can be modeled as a weighted combination of DOA kernels, $\mathbf{H}_{fs} = \sum_{o=1}^{O}\mathbf{W}_{fo}z_{so}$. The MU for the spatial weights relating direction to sources can be derived as

$$z_{so} \leftarrow z_{so}\frac{\sum_{k,f,t} b_{fks}g_{kts}\mathrm{tr}(\mathbf{X}_{ft}\mathbf{W}_{fo})}{\sum_{k,f,t} b_{fks}g_{kts}\mathrm{tr}(\hat{\mathbf{X}}_{ft}\mathbf{W}_{fo})}, \tag{16}$$

where $\mathbf{W}_{fo} = \mathbf{P}_{fo}\circ\mathbf{A}_o$. Opposite to the unconstrained SCM model in [20] and the DOA-based SCM models in [21] and [22], we propose to keep the phase matrix (here denoted as

PDCM or $\mathbf{P}_{fo}$) as a fixed parameter whereas the MU for the level covariance matrix $\mathbf{A}_o$ is expressed as

$$\mathbf{A}_o \leftarrow \mathbf{A}_o \frac{\sum_{s,f,t,k} z_{so} b_{fks} g_{kts} (\mathbf{X}_{ft} \circ \mathbf{P}^*_{fo})}{\sum_{s,f,t,k} z_{so} b_{fks} g_{kts} (\hat{\mathbf{X}}_{ft} \circ \mathbf{P}^*_{fo})}. \quad (17)$$

As in [20], after every iteration of the CNMF algorithm, some post-processing is required to make $\mathbf{A}_o$ Hermitian and positive semidefinite. In particular, we enforce the LDCM to be conjugate symmetric by using

$$\mathbf{A}_o \leftarrow \frac{1}{2}(\mathbf{A}_o + \mathbf{A}_o^H). \quad (18)$$

Then the eigenvalue decomposition is performed by $\mathbf{A}_o = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{H}}$ and the negative eigenvalues are set to zero, denoted as $\hat{\mathbf{D}}$. Finally, matrix $\mathbf{A}_o$ is enforced to be positive semidefinite by applying the following update:

$$\mathbf{A}_o \leftarrow \mathbf{U}\hat{\mathbf{D}}\mathbf{U}^{\mathrm{H}}. \quad (19)$$

Although no theoretical guarantee has been found, Eq. (17) together with the post-processing stage has demonstrated empirically to be always decreasing when the squared Euclidean distance is used.

B. Formulation for Itakura Saito Divergence

As explained in [20], the Itakura Saito divergence of the observed and estimated multichannel signal using the SCM representation can be expressed as

$$D_{IS}(\mathbf{X}_{ft}, \hat{\mathbf{X}}_{ft}) = \mathrm{tr}(\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}) - \log\det(\mathbf{X}_{ft}, \hat{\mathbf{X}}_{ft}^{-1}) - M. \quad (20)$$

Analogously to the Euclidean distance and omitting the constant terms, the cost function for IS divergence can be expressed as

$$f(\mathbf{A}, \mathbf{Z}, \mathbf{B}, \mathbf{G}) = \left[\mathrm{tr}(\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}) - \log\det(\hat{\mathbf{X}}_{ft})\right]. \quad (21)$$

Again, the optimization scheme of majorization proposed in [20] can be used (see Appendix B for details). Then, as in the case of the Euclidean distance, the update rules of b_{fks}, g_{kts}, z_{so} can be estimated for the IS divergence in Eq. (20) as

$$b_{fks} \leftarrow b_{fks} \sqrt{\frac{\sum_{t,o} z_{so} g_{kts} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}{\sum_{t,o} z_{so} g_{kts} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}} \quad (22)$$

$$g_{kts} \leftarrow g_{kts} \sqrt{\frac{\sum_{f,o} z_{so} b_{fks} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}{\sum_{f,o} z_{so} b_{fks} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}} \quad (23)$$

$$z_{so} \leftarrow z_{so} \sqrt{\frac{\sum_{f,t,k} b_{fks} g_{kts} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}{\sum_{f,t,k} b_{fks} g_{kts} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}}. \quad (24)$$

As in [20], update rules for the level matrix $\mathbf{A}_o$ are obtained by solving an algebraic Riccati equation

$$\mathbf{A}_o\mathbf{C}\mathbf{A}_o = \mathbf{D}, \quad (25)$$

where $\mathbf{C}$ and $\mathbf{D}$ are defined as

$$\mathbf{C} = \sum_{s,k} z_{so} b_{fks} \sum_{f,t} g_{kts} \hat{\mathbf{X}}_{ft}^{-1} \circ \mathbf{P}^*_{fo} \quad (26)$$

$$\mathbf{D} = \sum_{f} \Big(\mathbf{W}'_{fo}\Big(\sum_{s,k} z_{so} b_{fks} \sum_{t} g_{kts} \hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\Big)\mathbf{W}'_{fo}\Big) \circ \mathbf{P}^*_{fo}, \quad (27)$$

and $\mathbf{W}'_{fo}$ is the target matrix before the update. The solution to the Riccati equation is explained in Appendix C, where the obtained $\mathbf{A}_o$ after update is positive semidefinite (i.e. Eq. (19) is not required). Finally, to compensate for computer arithmetical error, Eq. (18) is used to ensure $\mathbf{A}_o$ to be Hermitian.

C. Parameter Scaling

Scaling the parameters is necessary to ensure that the SCM mixing filter only models the phase and the relative amplitude differences between channels. First of all, the scale of LDCM is constrained to ensure $||\mathbf{A}_o||_F = 1$ by applying

$$\mathbf{A}_o \leftarrow \frac{\mathbf{A}_o}{\mathrm{tr}(\mathbf{A}_o)}, \quad (28)$$

after every iteration of the CNMF algorithm for both IS and EUC cases. Then, in order to ensure numerical stability, the following scaling factors are applied to ensure that $\sum_o z_{so}^2 = 1$ and $\sum_l g_{kts}^2 = 1$.

$$\hat{b}_k = \left(\sum_{l=1}^{L} g_{kts}^2\right)^{\frac{1}{2}}, \quad g_{kts} \leftarrow \frac{g_{kts}}{\hat{b}_k}, \quad b_{fks} \leftarrow b_{fks}\hat{b}_k \quad (29)$$

$$\hat{a}_s = \left(\sum_{o=1}^{O} z_{so}^2\right)^{\frac{1}{2}}, \quad z_{so} \leftarrow \frac{z_{so}}{\hat{a}_s}, \quad b_{fks} \leftarrow b_{fks}\sum_s \hat{a}_s, \quad (30)$$

after the updates of z_{so} and g_{kts}, respectively, for both IS and EUC cases.

D. Incorporation the cross-correlation penalty terms

The penalty terms defined in Section III-B used to avoid overlapping directions between sources and restrict the weights to have a single predominant direction are incorporated to the cost function to be minimized:

$$D_*(\mathbf{X}_{ft}, \hat{\mathbf{X}}_{ft}) = D_*(\mathbf{X}_{ft}, \hat{\mathbf{X}}_{ft}) + \alpha_1\varphi_1(z_{so}) + \alpha_2\varphi_2(z_{so}), \quad (31)$$

where α_1 and α_2 are the parameters that control the importance of the regularized terms. In particular, the partial derivatives for the penalty terms φ_1 and φ_2 w.r.t. the spatial weights z_{so} are calculated as

$$\frac{\partial\varphi_1}{\partial z_{so}} = \frac{2\alpha_1}{S(S-1)}\varphi_1(z_{so}), \qquad \frac{\partial\varphi_2}{\partial z_{so}} = -\frac{2\alpha_2}{O(O-1)}\varphi_2(z_{so}). \quad (32)$$

TABLE I
SUMMARY OF THE OBSERVATIONS AND FIXED/FREE PARAMETERS FOR THE CNMF STAGES

	observation	fixed parameters	free parameters
Param. init & localiz.	$\angle\mathbf{X}_{ft}$	$\mathbf{P}_{fo}, \mathbf{A}_o$	z_{so}, b_{fks}, g_{kts}
Param. estimation	$\mathbf{X}_{ft}$	$\mathbf{P}_{fo}, z_{so}$	$\mathbf{A}_o, b_{fks}, g_{kts}$

Then, the update rule for parameter z_{so} for EUC distance is defined as

$$z_{so} \leftarrow z_{so} \frac{\sum\limits_{k,f,t} b_{fks} g_{kts} \mathrm{tr}(\mathbf{X}_{ft}\mathbf{W}_{fo}) + \frac{2\alpha_2}{O(O-1)}\varphi_2(z_{so})}{\sum\limits_{k,f,t} b_{fks} g_{kts} \mathrm{tr}(\hat{\mathbf{X}}_{ft}\mathbf{W}_{fo}) + \frac{2\alpha_1}{S(S-1)}\varphi_1(z_{so})}, \quad (33)$$

and for IS divergence as

$$z_{so} \leftarrow z_{so} \sqrt{\frac{\sum_{f,t,k} b_{fks} g_{kts} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo}) + \frac{2\alpha_2}{O(O-1)}\varphi_2(z_{so})}{\sum_{f,t,k} b_{fks} g_{kts} \mathrm{tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo}) + \frac{2\alpha_1}{S(S-1)}\varphi_1(z_{so})}}. \quad (34)$$

E. Algorithm Implementation

Due to the dimensionality of the model in Eq. (9) the algorithm to estimate the model parameters is divided into two stages. First, the parameters are initialized randomly and the spatial weights estimated accounting only to the phase differences between the channels. The observations are magnitude-invariant SCMs, consisting only of phase differences estimatedas

$$\angle\mathbf{X}_{ft} = \frac{\mathbf{X}_{ft}}{|\mathbf{X}_{ft}|}, \quad (35)$$

where $\mathbf{X}_{ft}$ is the signal SCM defined in Eq. (4). Note that a similar principle of estimating the locations of sources by using only phase information has been widely used in the literature (e.g. the SRP-PHAT algorithm [34]).

Second, after estimating the spatial weights z_{so}, we propose to estimate the level difference between channels (i.e. the LDCM $\mathbf{A}_o$) together with the other free parameters (i.e. b_{fks} and g_{kts}). Therefore, original signal SCM model in Eq. (4) is used as observation model.

The setup for both source localization and parameter estimation stages is presented in Table I. Finally, the whole proposed CNMF algorithm is detailed in Algorithm 1.

V. SOURCE RECONSTRUCTION

Once the model parameters have been optimized, we perform the source separation using generalized Wiener filtering. The estimated CNMF magnitude spectrogram for each sound source s can be defined from our proposed model in Eq. (6) as

$$\overline{y}_{msft} = \sum_o \mathrm{tr}(\mathbf{A}_o)_m z_{so} \sum_k b_{fks} g_{kts}. \quad (36)$$

Then the generalized Wiener mask is computed as

$$\tilde{\mathbf{y}}_{msft} = \tilde{\mathbf{x}}_{ft} \frac{\overline{y}_{msft}}{\sum_{s'o} \mathrm{tr}(\mathbf{A}_o)_m z_{s'o} \sum_k b_{fks} g_{kts}}. \quad (37)$$

Finally, the time-domain signals are obtained by inverse FFT and frames are combined by weighted overlap-add.

Algorithm 1 Pseudo code of the proposed CNMF algorithm

```
Initialize z_so, b_fks and g_kts with random values uniformly distributed between zero and one.
Initialize P_fo using Eq. (7) and [A_o]_nm = 1/M.
# PARAMETERS INITIALIZATION LOOP
Compute the input signal phase SCM using Eq. (35).
Compute the signal model using Eq. (6).
while not convergence and iter ≤ no. of iters do
    Update b_fks according to Eq. (14) (EUC) or (22) (IS)
    Recompute the signal model using Eq. (6).
    Update g_kts according to Eq. (15) (EUC) or (23) (IS)
    Scale g_kts to l2-norm and compensate by rescaling b_fks as specified in Eq. (29).
    Recompute the signal model using Eq. (6).
    Update z_so according to Eq. (33) (EUC) or (34) (IS).
    Scale z_so to l2-norm and compensate by rescaling b_fks as specified in Eq. (30).
    Recompute the signal model using Eq. (6).
end while
# PARAMETERS ESTIMATION LOOP
Compute the signal SCM observation using Eq. (4).
while not convergence and iter ≤ no. of iters do
    Update b_fks according to Eq. (14) (EUC) or (22) (IS).
    Recompute the signal model using Eq. (6).
    Update g_kts according to Eq. (15) (EUC) or (23) (IS).
    Scale g_kts to l2-norm and compensate by rescaling b_fks as specified in Eq. (29).
    Recompute the signal model using Eq. (6).
    Update A_o using Eq. (17) (EUC) or (26) (IS).
    Apply post-processing to enforce A_o to be hermitian using Eq. (18) (EUC and IS) and semipositive definite using Eq. (19) (only EUC).
    Scale A_o according to Eq. (28).
    Recompute the signal model using Eq. (6).
end while
```

VI. EVALUATION

A. Datasets

1) Two-channels recordings from SiSEC'08: The first dataset used in this work is a subset of the Signal Separation Evaluation Campaign (SiSEC 2008) [33] development dataset (dev1). In particular, subsets of synthetic convolutive mixtures and live recordings were used here. The dataset is composed of 40 mixture signals including:

- 32 speech signal mixtures (male and female).
- Four non-percussive music signal mixtures.
- Four music signal mixtures including drums.

The recordings were made using omnidirectional microphones. The room dimensions were 4.45 x 3.55 x 2.5 m. The reverberation time (T_{60}) was set to either 130 ms or 250 ms and the distance between the two microphones to either 5 cm or 1 m, resulting in four different configurations overall. The source directions of arrival varied between -60 degrees and $+60$ degrees with a minimal spacing of 15 degrees and the distances between the sources.

2) Four-channels recording conditions: A second dataset using two four-channel microphone arrays, with a small and

TABLE II
SOURCES SPATIAL POSITION PER MIXTURE IN THE GENERATED TWO AND THREE SOURCES FOUR-CHANNELS DATASET

	2 sources		3 sources		
mixture no.	1	2	1	2	3
1	45^o	90^o	0^o	45^o	90^o
2	135^o	180^o	45^o	90^o	135^o
3	0^o	90^o	0^o	45^o	-45^o
4	45^o	135^o	0^o	90^o	180^o
5	0^o	135^o	0^o	135^o	180^o
6	45^o	180^o	45^o	135^o	-45^o

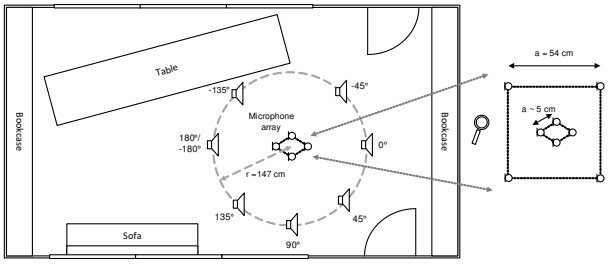

Fig. 3. Room, loudspeaker positions and microphone array placement.

large inter-microphone distance, were generated from impulse responses (IR) collected using the arrays in a regular meeting room.

The array with small microphone distances, introduced in [22], consists of four Sennheiser MKE-2 omnidirectional condenser microphones enclosed in a metal casing with approximate inter-microphone distance of 5 cm. A second larger array centered around the small array was used simultaneously to capture IRs. The larger array also consists four Sennheiser MKE-2 omnidirectional condenser microphones on the corners of square with side a = 54 cm. Hereafter the arrays are referred to as small and large array, respectively.

The recoding of IRs[2] was done in a room with dimensions of 7.95 m x 4.90 m x 3.25 m and having a average reverberation time of $T_{60} = 350$ ms. The measurement signal used was a MLS sequence of order 18, and Genelec G two loudspeaker was used to reproduce the measurement signal. The overview of the recording configuration and room layout is illustrated in Figure 3.

The anechoic material from (SiSEC'08) development dataset (dev1) was convolved with the obtained IRs resulting in six different mixtures for each set of sound signals. The different source spatial positions per mixture are detailed at Table II. In total, we generated 36 mixtures of two and 36 mixtures of three simultaneous sources for both arrays.

B. Evaluation Metrics

An objective evaluation of the performance of the separation method was done by comparing each separated signal to the spatial images of the original sources and using objective measures by BSS Eval toolbox [37], [38]. In fact, the use of objective measures based on energy ratios between the signal components, i.e., source to distortion ratio (SDR), the source to interference ratio (SIR), the source to artifacts ratio (SAR) and the source image spatial distortion ratio (ISR), has been the standard approach in the specialized scientific community to test the quality of extracted signals.

Moreover, the overall perceptual score (OPS), the target-related perceptual score (TPS), the interference-related perceptual score (IPS) and the artifacts-related perceptual score (APS) objective measures from the PEASS toolbox [39] were used with the aim of predicting the perceived quality of estimated source signals.

C. Algorithms for comparison

1) DBS+Wiener: A spatial signal processing field of beamforming [23] can be used for source separation. In particular, we have implemented a Delay and Sum Beamforming (DSB) design which consists of time aligning and summing the microphone signals. Knowing the array setup and the target DoA, i.e. beamformer look direction, DSB will enhance the sources originating from this direction. To perform a fair comparison with NMF-based source separation methods, a postprocessing Wiener filtering stage is applied to the output of DSB [24].
2) Multichannel NMF [25]: This method models the multichannel audio spectrogram using NMF with the Itakura Saito divergence. The method has two variants, instantaneous and convolutive mixing that are compared here. To estimate the mixing and source parameters, we have used the implementation provided by the authors using the expectation-maximization (EM) algorithm.
3) Spatial NTF [16]: This method is suitable only for the two channels scenario. In particular, the method implements a NTF approach using a spatially weighted parameter together with the ground-truth spatial cues (a.k.a source location prior information) to perform the separation. The method only accounts for amplitude difference between channels and thus, phase information is discarded.
4) SCM CNMF [19]: The multichannel SCM model without any directional constraints in Section II-C.
5) Baseline: The beamforming-inspired SCM model in [22] using the implementation provided by the authors, is used as the baseline for our experiments.
6) Proposed EUC: Our proposed SCM model using CNMF in Section III using the Euclidean distance for parameter estimation in Section IV.A.
7) Proposed IS: Our proposed SCM model using CNMF in Section III using the Itakura Saito divergence for parameter estimation in Section IV.B.

In the case of baseline and the proposed methods, three configurations are used depending on the initialization of the spatial weights parameter z_{so}: a) SRP-PHAT spatial cues (SRP), b) Ground-truth spatial cues (GT) and c) Random initialization. Note that when using spatial cues (SRP or GT) z_{so} is set to zero for all the directions above ± 10 degrees the spatial cue as in [22].

Additionally, to analyze the reliability of the evaluation measures in Section VI.B we have evaluated the separation performance using two extreme cases:

[2]The IRs and the code used for the four channels dataset together with some listening demos can be found at: http://www.cs.tut.fi/sgn/arg/IntensityCNMF/

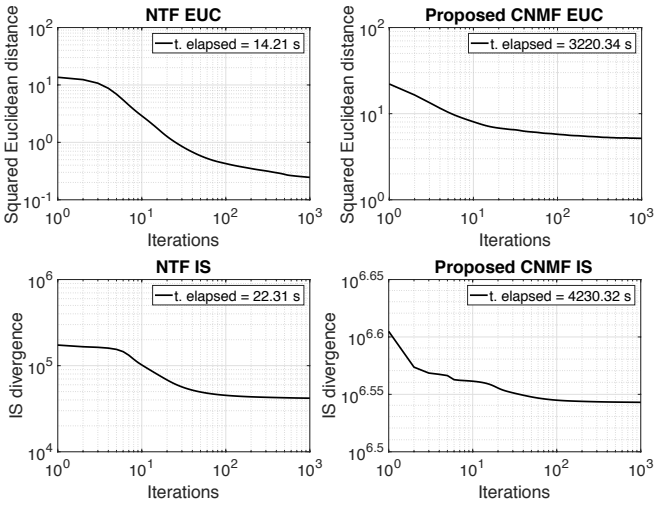

Fig. 4. Convergence behavior shown in log-log plots: EUC (top) and IS (bottom), NTF (left) and CNMF (right).

1) No separation: Using the mixture signal divided by the number of sources as input for the evaluation. This evaluation provides a starting point for the separation algorithms.
2) Random mask: The separated sources are obtained using the Wiener softmask strategy in (37) where the masks are generated using random values uniformly distributed between $[0, 1]$ but scaled to sum to unity.

D. Experimental Setup

In this paper, the time-frequency representation is obtained using 2048-point short-time Fourier transform (STFT) and half overlap between adjacent frames. Regarding the signal model, the parameters were set to similar values as used in related works [22], [16], and are as follows. The number of the basis functions $K = 30 \cdot S$, being S the number of sources in the mixture, and the maximum number of iterations for the decomposition is set to 300 for the localization loop and 500 for the parameters estimation loop.

The amount of look directions used with the SiSEC dataset was $O = 180$ which scanns the zero elevation plane. In case of four channel datasets, the full space around the arrays was considered in 7 elevations ($-67.5°$ to $67.5°$ with spacing of $22.5°$) and at each elevation 90 equally spaced azimuths was scanned (spacing of $4°$). The total amount of directions used in SCM modeling was 630. The parameters of the baseline can be found from [22]. For calculation of the localization constraints (10) and (11), the direction dependent weights at different elevations were summed to obtain a representation consisting of only azimuthal angles. The cross-correlation penalty could be extended to account for sources with different elevations as well, but for simplicity was left out from the paper.

E. Convergence Behavior

In Figure 4, the convergence behavior for 1000 iterations of the proposed SCM model using CNMF in Section III is compared with the NTF [16], [15] using both Euclidean distance and Itakura Saito divergence. The algorithms have been run for the SiSEC mixture "dev1_female3_liverec_130ms_1m", the signal duration is 30 seconds, the number of sources in the mixture S=3 and the number of components K is set 90. The computational time for each algorithm is displayed at the top right legend for each subfigure. We observe that the convergence behavior of the NTF algorithms is similar to that of the proposed SCM CNMF algorithms although the constant terms for the higher dimensionality SCM representation provoked a larger offset in the loss function for CNMF (specially for the proposed IS method), in comparison with the NTF algorithms. Regarding the computational time, CNMF is significantly slower than NMF algorithms and for both cases, IS divergence generally takes more time than EUC-NMF/CNMF.

TABLE III
COMPUTATIONAL TIME (IN SECONDS) FOR THE PROPOSED METHOD AND THE SOTA NMF-BASED METHODS IN SECTION VI-C

Methods / File	130ms1m	130ms5cm	250ms1m	250ms5cm
MultiNMF inst	427,86	492,01	503,70	466,42
MultiNMF conv	657,23	783,58	826,29	576,59
Spatial NTF	12,44	11,82	12,63	11,76
SCM CNMF	814,59	792,34	821,81	809,44
Baseline	5039,27	4980,83	5054,65	4915,97
Proposed EUC	1808,37	1789,13	2048,49	1903,53
Proposed IS	2940,22	2742,33	2886,81	2799,78

The computational time for all the NMF-based algorithms in section VI-C is presented in Table III. The algorithms were coded with Matlab and run on an Intel Xeon E5-2697 (2.60GHz) processor. For comparison, we have used the SiSEC "dev1_female3_liverec" with 30 seconds in length per file and four mixing conditions, 130 / 250 ms reverberation time and 1m/5cm microphone spacing (see Section VI.A). The setup is described in Section VI.D. As mentioned and due to its simplicity, NMF/NTF are faster than CNMF algorithms. In fact, the Spatial NTF in [16] using the MU algorithm is significantly faster than the other compared methods. The multichannel IS-NMF in [25] (MultiNMF inst) using the EM algorithm ranks second in terms of computational time.

For the CNMF algorithms, SCM methods using DoA kernels (Baseline and Proposed algorithm) are slower than the lower dimensionality SCM CNMF models in [19] and [25] (MultiNMF conv). Despite the number of iterations is higher (300 iters for localization and 500 for parameter estimation), the proposed method is faster than Baseline. In fact, the dimensionality of free parameters to be estimated during the localization is lower than in the parameter estimation procedure (see Table I). Moreover, once the localization is performed, the dimensions for the spatial weights and the PDCM and LDCM is reduced to a sparse set of looking directions by only retaining truly non-zero indices (see Figure 2).

F. Results with two-channels SiSEC development dataset

For the first evaluation, we have used the two-channels SiSEC'08 dev1 dataset described in section VI.A.1. The results are illustrated in Figures 5 and 6.

Since BSS is evaluated here, parameters for all the compared methods are randomly initialized except for the

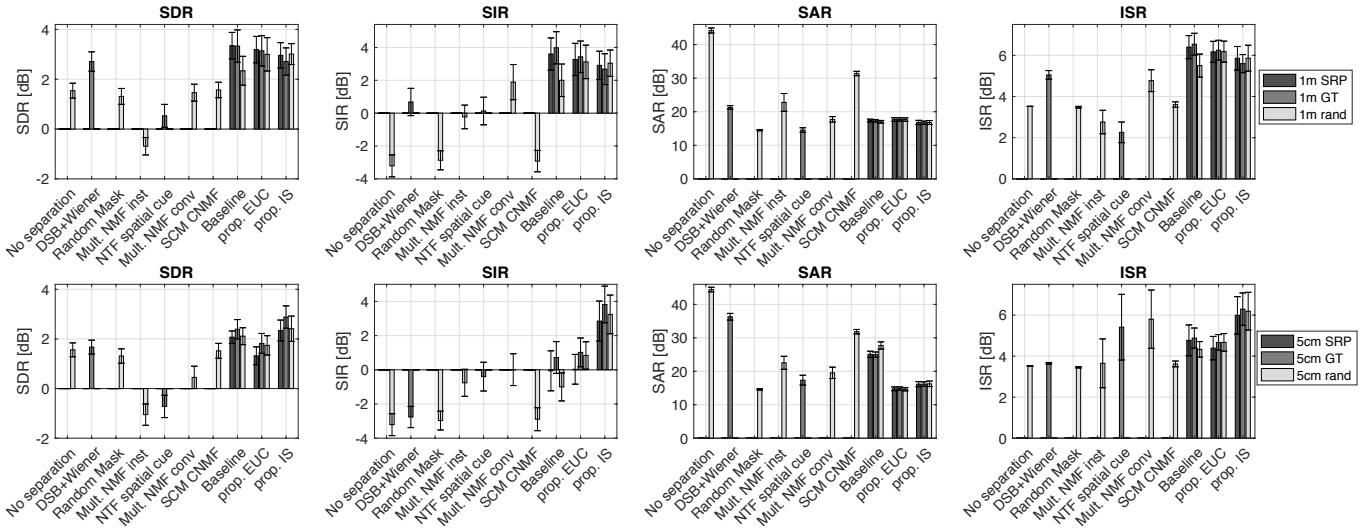

Fig. 5. Objective results using the BBS EVAL metrics [37], [38] for SiSEC development dataset, large array in the upper row and small array in lower row. Method abbreviated as (rand) uses blindly estimated DOA initialization whereas (GT) and (SRP) uses ground-truth and SRP-PHAT annotated DOA of sources, respectively. Vertical lines on top of each bar indicate 95% confidence intervals.

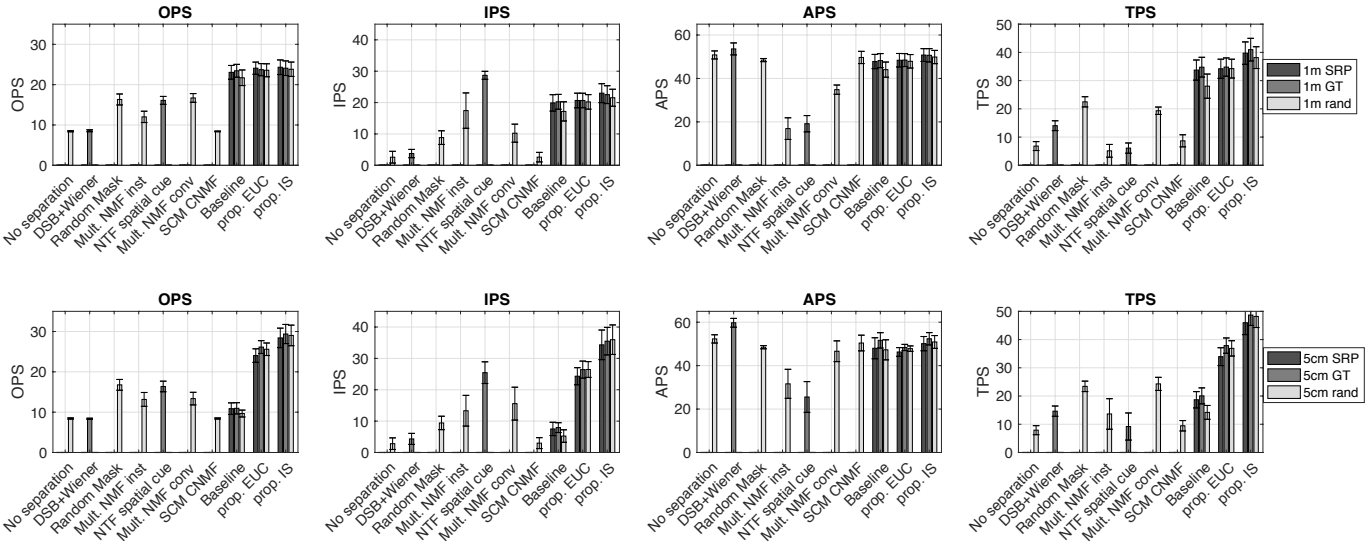

Fig. 6. Perceptual results using the PEASS metrics [39] for SiSEC development dataset, large array in the upper row and small array in lower row. Method abbreviated as (rand) uses blindly estimated DOA initialization whereas (GT) and (SRP) uses ground-truth and SRP-PHAT annotated DOA of sources, respectively. Vertical lines on top of each bar indicate 95% confidence intervals.

DSB+Wiener [24] and the NTF with spatial cues algorithm in [16] which are explicitly defined to be informed. Additionally, for the baseline and the proposed method, initialization of the spatial weight parameters using ground-truth and SRP-PHAT is presented together with the true BSS case (i.e. random initialization).

1) Results with 1 *m distance microphones:* We start by analyzing the values obtained with the extreme cases (i.e. no separation and random mask) in comparison with the other methods. Despite providing a poor subjective separation quality (examples provided online [3]), the obtained SDR values (see Figure 5) are very similar when compared to the conventional approaches, excluding the baseline and superior in terms of artifacts-related metrics (SAR, APS). Note that SAR values for all the methods are above 15 dB indicating good quality and the few higher SAR scores are obtained with very low SIR and no effective separation. Similar behavior can be observed for the PEASS metrics (see OPS and IPS values in Figure 6). For the sake of brevity, we propose to focus jointly on both interference-related (SIR, IPS) and overall metrics (SDR, OPS), as the main metrics in order to compare the different methods performance.

Despite its simplicity, the DSB+Wiener method provides better SDR than the compared methods except the baseline and the proposed approaches. However, this method suffers from the leakage of other sources into the extracted source resulting in a poor interference-related metrics (SIR, IPS) and the lowest OPS with respect to the proposed and baseline methods. Moreover, this method requires information about

[3]http://www.cs.tut.fi/sgn/arg/IntensityCNMF/

the DoAs for each source.

The multichannel NMF in [25] obtained the worst performance in terms of SDR across the compared methods (SDR= -0.69 dB, SIR= -0.23 dB). The NTF approach in [16] using information a priori (i.e. spatial cues) during the factorization allows to improve the separation performance w.r.t. the instantaneous model in [25] (here denoted as Mult. NMF inst) by 1 dB in SDR and 0.5 dB in SIR and obtains competitive results in terms of PEASS metrics. However, the mentioned methods do not take full advantage of the multichannel setup as they only use amplitude information to estimate the model parameters.

The SCM based models, i.e. SCM CNMF [19] and Mult. NMF conv [25] allow the estimation of the model parameters accounting for both amplitude and phase differences. However, without further constraints and especially when dealing with spectrally similar sources, these methods suffer to discriminate the sources in the mixture. The method in [25] clearly outperforms [19] in terms of interference-related metrics (SIR, IPS). Overall, similar performance is obtained in terms of SDR although the OPS higher for the method in [25]. Note that the CNMF-based method in [19] provides similar results than the extreme no separation case (see Figures 5 and 6).

The beamforming-based CNMF model [22] (here denoted as baseline) outperforms the models in [25], [19] as it relates phase differences with directions of arrivals and thus, allows discriminating the sources as a function of their spatial location. In fact, when spatial cues (i.e. ground-truth (GT) or SRP-PHAT (SRP) estimated source locations) are given a priori, this method obtains the most reliable results (SDR= 3.35 dB, SIR= 3.60 dB using GT and SDR= 3.33 dB, SIR= 3.98 dB with SRP). However, when no prior information is given (i.e. random spatial weights initialization), the method underperformed around 1 dB and 2 dB in terms of SDR and SIR, respectively. This underperformance can be observed also in the PEASS metrics. In fact, without further constraints, the model allows sources to have overlapping DOAs or to spread across multiple directions.

Regarding the proposed method, the obtained results are slightly below the baseline approach when spatial cues are given (SDR= 3.14 dB, SIR= 3.43 dB) using GT and (SDR= 3.19 dB, SIR= 3.27 dB) with SRP prior for the Euclidean distance. When using Itakura Saito as the objective function, the obtained results are (SDR= 2.71 dB, SIR= 2.68 dB) and (SDR= 2.96 dB, SIR= 2.91 dB) using GT and SRP prior, respectively. Looking at the PEASS metrics however, the situation is rather different, with the proposed method gaining the highest OPS and the IS optimization criterion slightly outperforming the EUC results.

Nonetheless, as pointed out in the earlier sections, the proposed algorithm does not require initialization of the spatial weights based on the DOA estimated prior the NMF model parameter estimation. This property is due to the twofold (localization-separation) strategy and the single/non-overlapping direction penalties of the spatial weights causing the algorithm to converge to spatially coherent yet discriminated solutions. In fact, for the completely blind case (i.e. randomly initialized spatial weights) the proposed method provides similar results to the informed case both BSSeval and PEASS metrics.

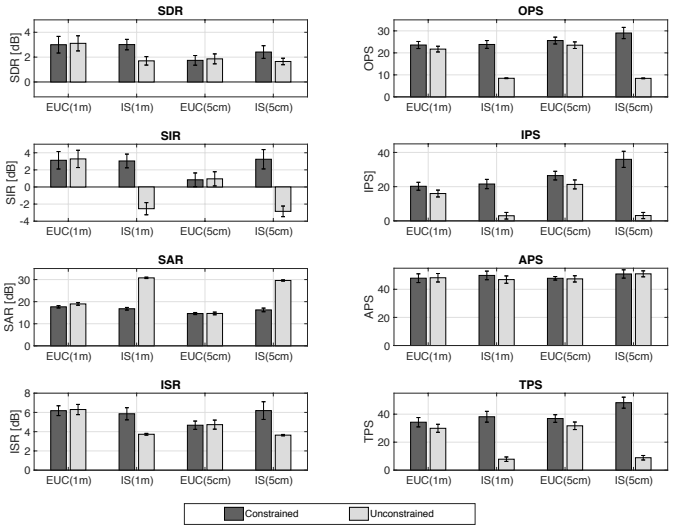

Fig. 7. Separation results using the BBS EVAL and the PEASS metrics for SiSEC development dataset of the proposed method using random initialization unconstrained (light bars) and using sparse localization constraints from Section III-B(dark bars). Vertical lines on top of each bar indicate 95% confidence intervals.

Note that the proposed method only uses phase information to estimate the spatial weights while the baseline method estimates all the parameters jointly using both amplitude and phase information. Although using amplitude information to estimate the directions of arrival of sources has been demonstrated to be beneficial for the case of large array microphone setups (i.e. baseline and the proposed method achieve similar performance with prior information), without prior information this method is prone to converge to unwanted local minima and thus, separation results are clearly degraded.

The effect of the localization constraints on the SiSEC dataset can be seen in Figure 7. In fact, under moderate reverberation conditions, the effect of the constraints for the proposed method using the EUC cost function is not significant (EUC relies more heavily on the largest coefficient). On the contrary, the effect of the single/non-overlapping directions penalties is more noticeable for the scale invariant IS cost function, improving significantly the separation results in terms of BSS and PEASS metrics (excluding the artifact-related ones).

2) Results with 5 *cm distance microphones:* For the case of 5 cm distance between microphones, it is interesting to see how the results of the real-valued NMF-based methods fall in comparison with the 1 m case as they rely only on the level difference between channels. In fact, when the microphones are close to each other, the level (or intensity) differences is practically negligible, and thus, time differences become the main spatial cue. In contrast, CNMF methods also take into account the phase differences and thus, provide superior separation results. Nonetheless, the lack of significant level difference between channels provokes an underperformance of the multichannel convolutive method in [25] by 1 dB in SDR and 2 dB in SIR w.r.t the 1m distance microphones. Otherwise,

the SCM CNMF method in [19] behaves similarly to the 1 m distance setup (SDR$= 1.52$ dB, SIR$= -2,89$ dB). In fact, the obtained results are in line with the extreme no separation case (in other words, no effective separation is performed). Similar performance is obtained by the DSB+Wiener method which underperforms mainly in terms of SIR in comparison with the 1 m setup. In general, PEASS metrics do not show significative differences to the compared methods (excluding the baseline and the proposed method) with respect to the 1 m distance setup.

The baseline approach in [22] underperformed when level differences between channels are negligible. In fact, separation results are around 1 dB below the SDR values obtained for the 1 m setup when spatial cues are given and 0.2 dB below the SDR in the non-informed (random) case. It is also interesting to see the baseline performance in terms of SIR (below 0.7 dB and -1 dB in the informed and random cases, respectively). In other words, without further constraints, jointly estimating the parameters using both amplitude and phase information provokes the method to converge to non-optimal solutions. Note that this poor separation performance is not noticeable if we only consider SDR, however, this low performance w.r.t the 1 m setup is more clear if we analyze the PEASS metrics (see Figure 6).

Regarding the proposed method, the strategy of estimating the spatial weights using only phase information and group sparsity constraints in a first place enforces the sparsity of the sources attending to their DOAs and provides a suitable prior to estimate the LDCM and the NMF parameters in the second stage. Consequently, better results are obtained, especially in terms of SIR, IPS and OPS. The scale-invariant IS clearly outperforms the EUC version in terms of interference-related metrics (3 dB in SIR and 11% in IPS). Using ground-truth spatial cues slightly improves the separation results in terms of BSSeval metrics w.r.t the non-informed (randomly initialized) case for the proposed method with IS by 0.4 dB in SDR and 0.6 dB in SIR while, for EUC distance the difference is not significant. On the contrary, using SRP-PHAT spatial cues slightly underperforms the blind case which demonstrates the robustness of our CNMF-based localization (i.e. spatial weights estimation) scheme.

G. Results with four-channels dataset

The result of evaluation with the four channel dataset are illustrated in Figures 8 and 10 for two simultaneous sources and Figures 9 and 11 for three simultaneous sources. In the following, we will concentrate on analysis of the benefits of the proposed model over baseline [22].

1) Results with large array: Similar to SiSEC'08 dataset with 1 m microphone distance, the beamforming-based CNMF model [22] and the proposed method perform equally with the large array (54 cm) when using prior information for initialization. For the two-sources dataset the SDR is around 7 dB and SIR is around 10 dB for the baseline and proposed method with EUC as cost function. The separation performance decreases to $4-5$ dB in SDR and $6-7$ dB in SIR with three simultaneous sources. The IS cost function for the proposed method receives slightly lower separation performance than EUC in both datasets in terms of overall (SDR, OPS) and interference-related metrics (SIR, IPS). The SCM CNMF method [19] achieved very modest performance in comparison to the proposed method but results in less artifacts across the compared methods (similar to the two channels experiment).

The performance of delay and sum beamforming with postfiltering (DSB+Wiener) benefits from the higher number of microphones and provides very competitive results in terms of BSSeval metrics ($3.8-6.1$ dB in SDR and $3.2-7$ dB in SIR). Regarding the PEASS metrics, this method clearly underperforms the baseline and the proposed methods in terms of OPS, IPS and TPS but slightly outperforms the SCM CNMF method [19] except in terms of artifact-related metrics (SAR, APS).

The performance of the baseline with random initialization is relatively improved in comparison to the SiSEC dataset and it achieves the same or even improved performance in comparison to initialization with SRP estimated source locations. This can be argued to be caused by having less spatial ambiguity due to increased amount of microphones. Moreover, under higher reverberant conditions ($T_{60} = 350$ ms for the four channels dataset while for two channels is $130-250$ ms), restricting the direction weights to a single DOA per source may cause a poor modelling of the reflections in comparison with the unconstrained baseline model. Nonetheless, with four channels and two sources the absolute performance obtained with the proposed method (SDR 7.7 dB and SIR 11.2 dB) is highest among all datasets and conditions, which indicates that the separation performance in terms of BSSeval metrics scales according to the difficulty of the problem, i.e. the amount of sources to be separated and the number of microphones available given that sufficient level differences exist for the proposed model. Regarding the PEASS metrics, the proposed method demonstrated to be robust against the number of sources / conditions providing the best averaged results in terms of OPS, IPS and TPS.

2) Results with small array: The performance of the proposed method with respect to baseline and SCM CNMF method [19] with the small array is very similar in comparison to the SiSEC dataset with 5 cm microphone distance. The proposed method using the IS optimization criterium achieves the best results in terms of interference-related metrics (SIR, IPS). On average the separation performance is around 4.5 dB and 2.5 dB in SDR and 9 dB and 4.5 dB in SIR for two and three sources datasets, respectively. Additionally, noticeably better OPS and IPS scores are obtained with the proposed method than with DSB+Wiener, baseline or SCM CNMF method [19].

The performance of the baseline with random initialization is greatly reduced (by 1.5 dB and 1.0 dB in SDR and 1.2 dB and 2.3 dB in SIR), whereas the use of prior information and completely random initialization with the proposed method has only small effect on separation performance. It can be observed from the results of all datasets that the proposed method with IS as cost function is favored with small arrays in terms of separation performance, whereas with larger arrays the EUC achieves slightly better results.

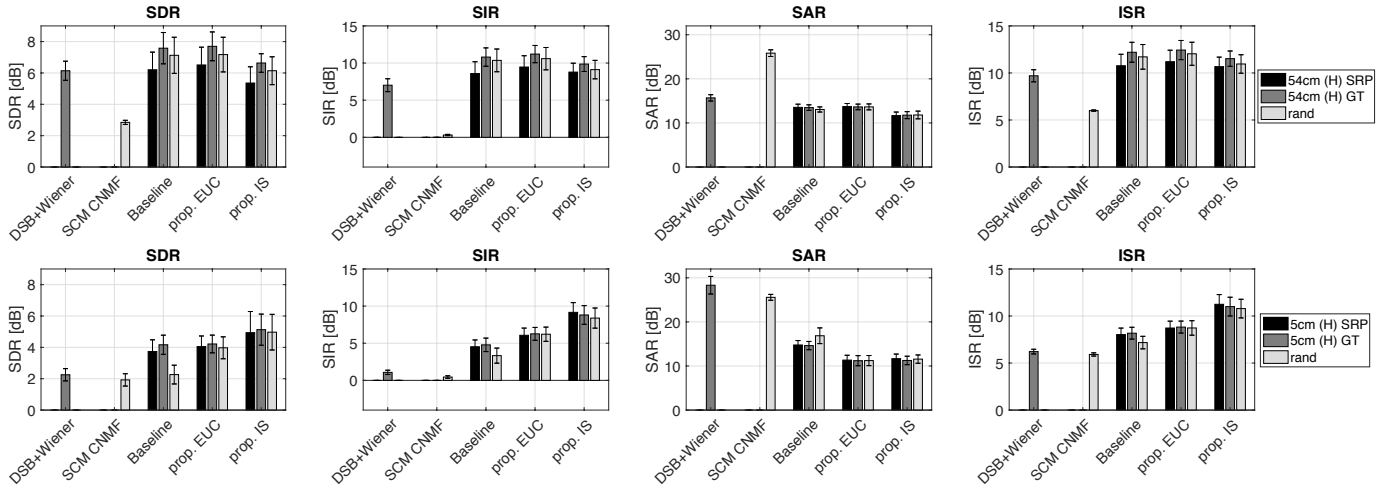

Fig. 8. Results with four channel arrays with two simultaneous sources, large array in the upper row and small array in lower row. Vertical lines on top of each bar indicate 95% confidence intervals.

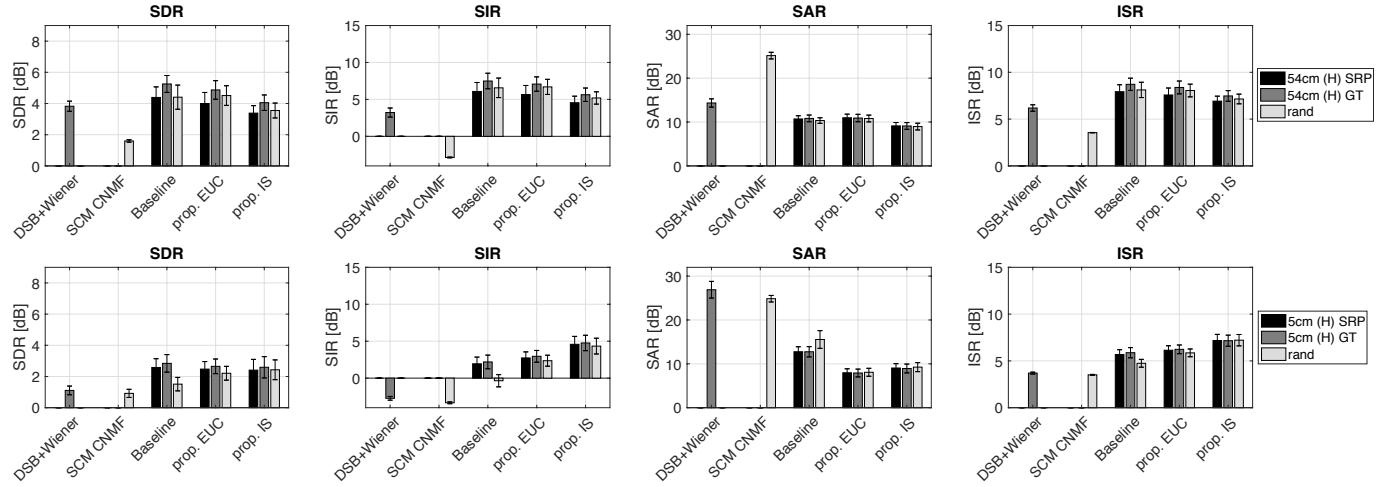

Fig. 9. Results with four channel arrays with three simultaneous sources, large array in the upper row and small array in lower row. Vertical lines on top of each bar indicate 95% confidence intervals.

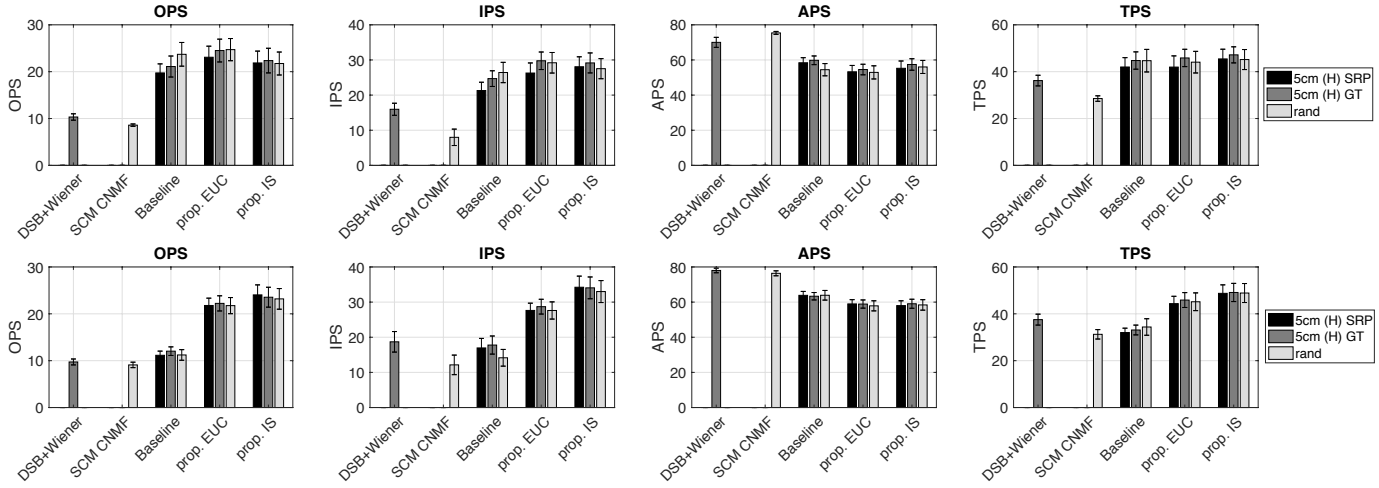

Fig. 10. Results using the PEASS metrics [39] with four channel arrays with two simultaneous sources, large array in the upper row and small array in lower row. Vertical lines on top of each bar indicate 95% confidence intervals.

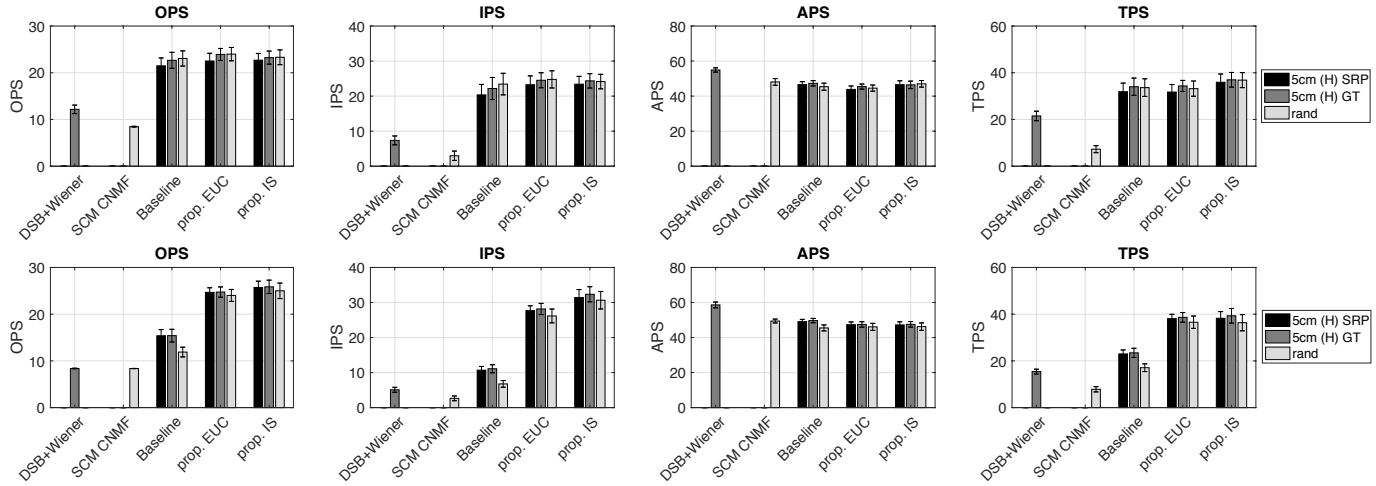

Fig. 11. Results using the PEASS metrics [39] with four channel arrays with three simultaneous sources, large array in the upper row and small array in lower row. Vertical lines on top of each bar indicate 95% confidence intervals.

H. Discussion

Separation performance of the proposed method using level and time difference SCMs was evaluated using two-channel SiSEC'08 development dataset and four-channels signals recorded in a regular room with moderate reverberation.

First, the separation performance has been compared with other NMF/CNMF state-of-the-art methods together with two extreme scenarios (no separation and random mask) and a classical beamforming technique using a postfiltering for the case of the two-channel dataset. The proposed method clearly improves the separation results w.r.t the compared real-valued NMF and non-beamforming based CNMF methods except for the case of artifact-related metrics (SAR, APS) where all the compared methods obtain high scores. When spatial cues are given, the results of baseline and the proposed method in terms of BSSeval metrics are comparable. However, in the true blind case, the proposed method behaves similarly to the informed case and thus outperforms the baseline, especially in terms of SIR for the short distance microphone setup. In fact, the proposed twofold (localization-separation) strategy and the group sparsity constraints imposed to the spatial weights make our algorithm more robust against different microphone setups than the other compared methods. This outperformance is statistically significant in terms of PEASS metrics (OPS, IPS and TPS) for the short distance microphone setup (see 95% CIs in Figure 6).

Then, we evaluate the performance of the proposed method w.r.t. the baseline, the SCM-CNMF in [20] and the DSB with postfiltering using the four-channels dataset. In general, the obtained separation results for all the compared methods scales according to the difficulty of the problem, i.e. the amount of sources to be separated and the number of microphones available. Nonetheless, the proposed method demonstrated to be robust against the number of sources and conditions providing the best averaged results in terms of the PEASS OPS, IPS and TPS metrics (as for the two channel dataset, differences are statistically significant for the small array setup). Besides, using the Euclidean distance as the optimization criterion improves the results for large spacing arrays whereas the scale-invariance property of the Itakura Saito slightly improves the results for the case of short spacing arrays. Because of the space limitation, further analysis of the results as a function of the source type has not been included in this paper. In fact, with large distance arrays, the proposed signals perform better for speech than music signals whereas for short distance microphones results are comparable. In general, results are degraded for music mixtures with percussive signals.

VII. Conclusions

In this paper we proposed a SCM model of the mixture complex-valued spectrogram that uses level and time differences for the task of multichannel sound source separation. These level and time differences are modeled as the weighted sum of direction of arrival (DOA) kernels spanned across every possible source direction. In order to estimate the model parameters, a CNMF algorithm is proposed using both Euclidean and Itakura Saito optimization functions. Opposite to other SCM models in the literature, the source localization is implicitly defined in the model and two cross-correlation (i.e. group sparsity) constraints are imposed to the spatial weights parameter to enforce a single direction per source and no overlapping between source directions. We have shown the robustness of the proposed algorithm in comparison with other state-of-the-art BSS methods using various types of microphone array setups. In the future, we would investigate the suitability of the localization scheme for source localization problems. Additionally we will investigate the effect of the reverberation time for the proposed model.

Appendix A
Derivation of the Algorithm for EUC Distance

To minimize function $f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G})$ in Eq. (13), we follow the optimization scheme of majorization as in [20]. First, the following auxiliary function is defined:

$$f^{+}(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G},\mathbf{R}) = \sum_{f,t,s,o,k} b_{fks}^2 g_{kts}^2 z_{so}^2 \operatorname{tr}\left(\mathbf{W}_{fo}\mathbf{R}_{ftsok}^{-1}\mathbf{W}_{fo}^H\right) - \sum_{f,t,s,o,k} b_{fks} g_{kts} z_{so} \operatorname{tr}\left(\mathbf{X}_{ft}\mathbf{W}_{fo}^H\right) - \sum_{f,t,s,o,k} b_{fks} g_{kts} z_{so} \operatorname{tr}\left(\mathbf{W}_{fo}\mathbf{X}_{ft}^H\right), \quad (38)$$

where $\mathbf{R}_{ftsok}$ is Hermitian positive definite and satisfies $\sum_{s,o,k} \mathbf{R}_{ftsok} = \mathbf{I}^{M\times M}$, which is $M \times M$ identity matrix. As demonstrated in [20], the auxiliary function satisfies the two following conditions:

1) $f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G}) \leq f^{+}(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G},\mathbf{R})$
2) $f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G}) = min_R f^{+}(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G},\mathbf{R})$

That is, minimization of f^{+} with respect to $\mathbf{A}$, $\mathbf{Z}$, $\mathbf{B}$ and $\mathbf{G}$ can be used for an indirect optimization of f. Note that minimization of f implies the optimization of the model parameters with respect to Eq. (12). In fact, the equality $f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G}) = f^{+}(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G},\mathbf{R})$ is achieved by defining the auxiliary variable $\mathbf{R}_{ftsok}$ as

$$\mathbf{R}_{ftsok} = \hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo}z_{so}b_{fks}g_{kts}. \tag{39}$$

Therefore, minimization of function f is performed in two steps:

1) Minimize f^{+} with respect to $\mathbf{R}_{ftsok}$ using Eq. (39) to $f = f^{+}$.
2) Minimize f^{+} with respect to $\mathbf{A}$, $\mathbf{Z}$, $\mathbf{B}$ and $\mathbf{G}$

Then, as in [20], the derivation of the algorithm updates is achieved via partial derivation of function f^{+} w.r.t each model parameter and setting these derivatives to zero. However, for the sake of brevity, details about the computation of the partial derivatives are omitted here.

APPENDIX B
DERIVATION OF THE ALGORITHM FOR IS DIVERGENCE

Similar to the case of the EUC distance, the function $f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G})$ in Eq. (21) is minimized using the optimization scheme of majorization from [20]. In particular, the auxiliary function for the case of IS is defined as

$$f^{+}(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G},\mathbf{R},\mathbf{U}) = \sum_{i,l}\left[\log\det\mathbf{U}_{ft} + \frac{\det\mathbf{X}_{ft} - \det\mathbf{U}_{ft}}{\det\mathbf{U}_{ft}} + \sum_{s,o,k}\frac{\mathrm{tr}(\mathbf{X}_{ft}\mathbf{R}_{ftsok}^{H}\mathbf{W}_{fo}^{-1}\mathbf{R}_{ftsok})}{z_{ok}b_{fks}g_{kts}}\right], \tag{40}$$

where auxiliary variables $\mathbf{R}_{ftsok}$ and $\mathbf{U}_{ft}$ are hermitian positive definite matrices satisfying $\sum_{s,o,k}\mathbf{R}_{ftsok} = \mathbf{I}^{M\times M}$ and $\mathbf{U}_{ft} = \mathbf{U}_{ft}^{H}$. As for the Euclidean distance, the auxiliary function f^{+} fulfill the following conditions:

1) $f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G}) \leq f^{+}(\mathbf{A},\mathbf{Z},\mathbf{G},\mathbf{G},\mathbf{R},\mathbf{U})$
2) $f(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G}) = \mathbf{min}_{\mathbf{R},\mathbf{U}} f^{+}(\mathbf{A},\mathbf{Z},\mathbf{B},\mathbf{G},\mathbf{R},\mathbf{U})$,

and $f = f^{+}$ when

$$\mathbf{R}_{ftsok} = \hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo}z_{so}b_{fks}g_{kts}, \qquad \mathbf{U}_{ft} = \hat{\mathbf{X}}_{ft}. \tag{41}$$

Similar to the Euclidean distance, minimization of function f is performed in two steps:

1) Minimize f^{+} with respect to $\mathbf{R}_{ftsok}$ and $\mathbf{U}_{ft}$ using Eq. (41) to $f = f^{+}$.
2) Minimize f^{+} with respect to $\mathbf{A}$,$\mathbf{Z}$,$\mathbf{B}$ and $\mathbf{G}$.

Finally, multiplicative update rules are obtained after the second step, that is, after performing the partial derivatives of f^{+} w.r.t each parameter and equalizing to zero.

APPENDIX C
SOLVING AN ALGEBRAIC RICCATI EQUATION

As proposed in [19], the solution of the Riccati equation in Eq. (26) is detailed here. First, lets define a $2M \times 2M$ matrix $\mathbf{E}$ as

$$\mathbf{E} = \begin{bmatrix} 0 & -\mathbf{C} \\ -\mathbf{D} & 0 \end{bmatrix}. \tag{42}$$

We compute the $2M$ eigenvectors from $\mathbf{E}$ as $\mathbf{e_1}, ..., \mathbf{e_{2M}}$. Then, we sort the eigenvectors according to the associated eigenvalues in ascending order and remove the vectors corresponding to the smallest eigenvalues which in theory (i.e. omitting computer arithmetical errors) discards the negative eigenvalues. As a result, we will have M sorted eigenvectors $\mathbf{e'_1}, ..., \mathbf{e'_M}$.

Then, the new $\mathbf{A}_o$ is obtained as $\mathbf{A}_o \leftarrow \mathbf{I}\mathbf{J}^{-1}$, where the $M \times M$ matrices $\mathbf{J}$ and $\mathbf{I}$ are defined from the sorted eigenvectors $\mathbf{e'}$ as $\mathbf{J} = [e'_{1,1:M}, ..., e'_{M,1:M}]$ and $\mathbf{I} = [e'_{1,M+1:2M}, ..., e'_{M,M+1:2M}]$.

ACKNOWLEDGMENT

We wish to thank the reviewers for many very helpful comments and suggestions.

REFERENCES

[1] J. Woodruff, B. Pardo, and R. B. Dannenberg, Remixing stereo music with score-informed source separation, in Proc. Int. Conf. Music Information Retrieval (ISMIR) 2006, pp. 314-319.
[2] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing (ICASSP) 2012, pp. 57-60.
[3] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. Okuno, Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models, in Proc. Int. Conf. Music Information Retrieval (ISMIR), 2008, pp. 133-138.
[4] S. Ewert and M. Mller, Estimating note intensities in music recordings, in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP) 2011, pp. 385-388.
[5] J. F. Cardoso, "Blind signal separation: Statistical principles," in Proceedings of the IEEE, vol. 86, no. 10. IEEE Computer Society Press, October 1998, pp. 2009-2025.
[6] M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP) 2002, pp. 881-884.
[7] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," IEEE Trans. Audio, Speech, Language Process 2007, 15(5):1592-1604.
[8] F. Nesta, M. Omologo, and P. Svaizer, "Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS," in IEEE Workshop on Machine Learning for Signal Processing 2008, pp. 43-48.
[9] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), vol. 5, 2000, pp. 2985-2988.
[10] H. Sawada, S. Araki, and S. Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, IEEE Trans. Audio, Speech, Language Process, vol. 19, no. 3, pp. 516-527, 2011.
[11] J. Thiemann and E. Vincent, "An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement," in Proc. IEEE Int.Workshop Mach. Learning Signal Process., pp. 1–5, 2013.
[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov. "A consolidated perspective on multimicrophone speech enhancement and source separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4):692–730, 2017.

[13] Parry RM, Essa IA: Estimating the spatial position of spectral components in audio. In Proceedings of the 6th International Conference of Independent Component Analysis and Blind Signal Separation (ICA06). Charleston; March 2006:666-673.
[14] FitzGerald D, Cranitch M, Coyle E: Non-negative tensor factorisation for sound source separation. In Proceedings of the Irish Signals and Systems Conference. September 2005:8-12.
[15] Fevotte C, Ozerov A: Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues. In Proceedings of the 7th International Symposium on Computer Music Modeling and Retrieval (CMMR), 2010:102-115.
[16] Mitsufuji Y, Roebel A: "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2013, pp: 71-75
[17] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2009, pp. 3437-3440.
[18] F. J. Rodriguez-Serrano, S. Ewert, P. Vera-Candeas and M. Sandler, "A Score-Informed Shift-Invariant Extension of Complex Matrix Factorization for Improving the Separation of Overlapped Partials in Music Recordings," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2016, pp. 61-65.
[19] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel nmf," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2011, pp. 153156.
[20] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data," IEEE Trans. Audio, Speech, Language Process, vol. 21, np. 5, 2013
[21] J. Nikunen and T. Virtanen. "Direction of Arrival Based Spatial Covariance Model for Blind Sound Source Separation," IEEE Trans. Audio, Speech, Language Process, Volume 22, Issue 3, pp. 727 - 739, 2014.
[22] J. Nikunen and T. Virtanen, "Multichannel audio separation by Direction of Arrival Based Spatial Covariance Model and Non-negative Matrix Factorization," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 2014
[23] I. Tashev, "Sound capture and processing: practical approaches," John Wiley & Sons Inc, 2009.
[24] C. Marro, Y. Mahieux and K.U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques based on Microphone Arrays with Postfiltering", IEEE Trans. On Speech and Audio Processing, vol. 6, pp. 240-259, 1998.
[25] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. Audio, Speech, Language Process, vol. 18, no. 3, pp. 550-563, 2010.
[26] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proceedings of International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 1-4.
[27] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *European Signal Processing Conference (EUSIPCO)*, 2016.
[28] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," IEEE Trans. Audio, Speech, Language Process, vol. 24, no. 10, pp. 1652-1664, 2016.
[29] D. Malioutov, M. Cetin and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," in IEEE Transactions on Signal Processing, vol. 53, no. 8, pp. 3010-3022, Aug. 2005.
[30] J. Yin and T. Chen, "Direction-of-Arrival Estimation Using a Sparse Representation of Array Covariance Vectors," in IEEE Transactions on Signal Processing, vol. 59, no. 9, pp. 4489-4493, Sept. 2011.
[31] N. Murata, S. Koyama, H. Kameoka, N. Takamune, and H. Saruwatari, "Sparse sound field decomposition with multichannel extension of complex NMF," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 395–399, Shanghai, 2016.
[32] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," Neural Computation, vol. 21, no. 3, pp. 793?830, 2009.
[33] E. Vincent, S. Araki, F.J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B.V. Gowreesunker, D. Lutter and N.Q.K. Duong, The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges (external link), Signal Processing, 92, pp. 1928-1936, 2012.
[34] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Microphone arrays: signal processing techniques and applications," Eds: Michael Brandstein and Darren Ward, Springer-Verlag, 2001.
[35] S. Raczynski, N. Ono, S. Sagayama, Multipitch analysis with harmonic nonnegative matrix approximation, in Proc. Int. Conf. Music Information Retrieval (ISMIR) 2007, pp. 381386.
[36] F.J. Canadas-Quesada, P. Vera-Candeas, D. Martinez-Munoz, N. Ruiz-Reyes, J.J. Carabias-Orti, P. Cabanas-Molero, Constrained non-negative matrix factorization for score-informed piano music restoration, Digital Signal Processing, Volume 50, Pages 240-257, 2016.
[37] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech, Language Process, vol. 14, no. 4, pp. 1462-1469, 2006.
[38] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," Independent Component Analysis and Signal Separation, pp. 552-559, 2007.
[39] V Emiya, E Vincent, N Harlander, V Hohmann, "Subjective and objective quality assessment of audio source separation," IEEE Trans. Audio, Speech, Language Process, 19(7), 2046-2057, 2011

Julio Carabias received the M.Sc. degree in computer science and the Doctor of Science degree from the University of Jaen, Spain, in 2006 and 2011, respectively. He is currently a Postdoctoral Researcher with the Telecommunication Engineering Department at University of Jaen, Spain. His research interests are signal processing and machine learning. In particular, with a focus on signal decomposition methods for music signal processing applications including music transcription, source separation and audio to score alignment.

Joonas Nikunen received the M.Sc. degree in signal processing and communications engineering and the Ph.D. degree in signal processing from the Tampere University of Technology (TUT), Tampere, Finland, in 2010 and 2015, respectively. He is currently a Postdoctoral Researcher with TUT focusing on sound source separation with applications on spatial audio analysis, modification, and synthesis. His other research interests include microphone array signal processing, 3-D/360 audio in general and machine, and deep learning for source separation.

Tuomas Virtanen received the M.Sc. and Doctor of Science degrees in information technology from the Tampere University of Technology (TUT), Tampere, Finland, in 2001 and 2006, respectively. He is currently a Professor with the Laboratory of Signal Processing, TUT. He is known for his pioneering work on single-channel sound source separation using nonnegative matrix factorization-based techniques, and their application to noise-robust speech recognition, music content analysis, and sound event detection. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored about 170 scientific publications on the above topics. He is a member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE Signal Processing Society.

Pedro Vera was born in Madrid, Spain, in 1976. He received the M.S. degree in Telecommunication Engineering from the University of Malaga (UMA) in 2000 and the Ph.D. degree from the University of Alcala in 2006. Since 2000, he has worked at the Telecommunication Engineering Department of the University of Jaen. Nowadays, he is an Associate Professor in Signal Processing and Communications Area. His areas of research interest are Signal Processing and its Applications to Audio Analysis and NDT. He has been involved in research projects of the Spanish Ministry of Science and Education (MEC) and private companies.