Direction of Arrival Based Spatial Covariance Model for Blind Sound Source Separation

Joonas Nikunen, *non-Member* and Tuomas Virtanen, *Member, IEEE*

***Abstract*—This paper addresses the problem of sound source separation from a multichannel microphone array capture via estimation of source spatial covariance matrix (SCM) of a short-time Fourier transformed mixture signal. In many conventional audio separation algorithms the source mixing parameter estimation is done separately for each frequency thus making them prone to errors and leading to suboptimal source estimates. In this paper we propose a SCM model which consists of a weighted sum of direction of arrival (DoA) kernels and estimate only the weights dependent on the source directions. In the proposed algorithm, the spatial properties of the sources become jointly optimized over all frequencies, leading to more coherent source estimates and mitigating the effect of spatial aliasing at high frequencies. The proposed SCM model is combined with a linear model for magnitudes and the parameter estimation is formulated in a complex-valued non-negative matrix factorization (CNMF) framework. Simulations consist of recordings done with a hand-held device sized array having multiple microphones embedded inside the device casing. Separation quality of the proposed algorithm is shown to exceed the performance of existing state of the art separation methods with two sources when evaluated by objective separation quality metrics.**

***Index Terms*—multichannel source separation, spatial covariance models, non-negative matrix factorization, direction of arrival estimation, array signal processing**

I. Introduction

When recording an auditory scene using one or multiple microphones, it is preferred that the sound source dependent information can be separated for the uses of great variety of subsequent audio processing tasks. The examples of such applications include spatial audio coding (SAC) [1], [2], 3D sound analysis and synthesis [3], and signal enhancement for various purposes, such as automatic speech recognition (ASR) [4], [5]. When no prior information of the sources involved in the capture is available, the process is called blind source separation (BSS). The BSS problem in the case of spatial audio captures consist of decomposing the multichannel mixture signal into source signals and representing information about their spatial position or response from their originating location to each receiving microphone.

A well known BSS approach is the independent component analysis (ICA) [6] applied separately at each frequency of a short-time Fourier transformed (STFT) array input. It leads to an arbitrary frequency-wise source ordering referred to as the permutation problem. Source permutation is usually solved based on time difference of arrival (TDoA) interpretation of ICA mixing parameters [7]–[9]. The TDoAs calculated from phase differences become ambiguous when the frequency exceeds the spatial aliasing limit, which corresponds to a wavelength greater than half of the microphone spacing. As a result, the TDoAs cannot be directly utilized in solving the permutation problem for high frequencies. Additionally, the ICA parameters for a single source concatenated over frequency do not explicitly have a connection to the spatial position of a source but only to the phase difference caused by it. Separation methods directly utilizing TDoAs between microphones and creating time-frequency separation masks by clustering the measured TDoAs at each frequency include for example DUET [10] and binwise clustering [11].

More recently, methods based on finding spectrally redundant parts by non-negative matrix factorization (NMF) [12]–[14] have been proposed for separation of sound sources both with single [15], [16] and multichannel mixtures [17]–[20]. NMF is applied in the magnitude spectrogram domain and it finds an approximation of the mixture spectrogram using a linear combination of components that have a fixed spectrum and time-dependent gain. In the NMF separation framework the spatial properties of the sources can be modeled using a spatial covariance matrix (SCM) for each source at each STFT frequency bin [18]–[22]. Such extensions are hereafter referred to as complex-valued NMF (CNMF). The SCM denotes the mixing of the sources by magnitude and phase differences between the recorded channels, and is not dependent on the absolute phase of the source signal. Additionally, non-negative tensor factorization with spatial cues based on the magnitude panning of sources have been proposed in [23].

The CNMF algorithms [19]–[21] estimate unconstrained SCMs at each frequency, thus relying on the ability of the NMF magnitude model to separate repetitive parts that correspond to sources at a single spatial location. In the case of spectrally similar sources, for example two speakers, a single NMF component and the corresponding SCM can end up representing both the sources at different spatial locations. In such case, the estimated parameters cannot provide separation of the two sources. Spatial aliasing makes the algorithm prone to SCM estimation errors and separation is dependent on the magnitude model separation abilities. The CNMF method proposed in [18] assigns a fixed number of NMF components per source but it is reported to have a poor separation quality without an oracle initialization of the source parameters.

A spatial signal processing field of beamforming [24] can also be considered as a separation technique. The simplest design is the delay and sum beamformer (DSBF), which consists of time aligning and summing the microphone signals.

J. Nikunen and T. Virtanen are with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, email: firstname.lastname@tut.fi

The alignment correspond to a time delay caused by the target DoA, i.e. beamformer look direction, and the sources originating from this direction become enhanced. Other types of beamformers, such as the minimum variance distortionless (MVDR) beamformer, aim at suppressing and canceling interfering signals originating from other than the beamformer look direction. More recent advances in beamforming are based on adaptively estimating the noise characteristics and designing blocking matrix for the general sidelobe canceller structure [25], [26].

Beamforming methods assume that the geometry of the array is known and require a high number of microphones to work efficiently, to form a narrow beam that is useful for source separation. This is conventionally only achieved with physically large arrays. Beamformers also suffer from spatial aliasing which causes signal amplification from undesired source directions, i.e. peaks in the sidelobe structure. An emerging approach for beamforming based source separation is spherical array beamforming [27], [28] where a high number of densely spaced sensors in a physically small sphere is used to obtain uniform directivity properties towards every look direction.

In this paper we propose a novel BSS method that combines SCM estimation by beamforming-inspired DoA kernels and object-based signal analysis using NMF. We propose to model the source SCMs as the weighted combination of DoA kernels and the magnitudes of sources by the NMF. The DoA kernels represent the phase difference between array channels caused by a single source TDoA at a certain spatial position. The main benefit of the method comes from making the connection between the SCMs at each frequency by representing SCMs over frequency as a weighted sum of DoA kernels thus avoiding source and frequency ambiguity issues. By only estimating direction-dependent weights for the DoA kernels of each source, the proposed SCM model is jointly optimized over all frequencies, and produces better estimates of the source SCMs. The direction weights estimation also mitigates the effect of spatial aliasing in high frequencies due the estimation algorithm taking into account phase difference evidence across frequency by single time delays of individual DoA kernels.

The NMF components represent parts of the sources by estimating repetitive magnitude structures from the mixture signal. Components sharing the same spatial position are assumed to originate from the same source and can be thus linked together by simple clustering applied on the estimated direction weights. In addition to doing separation of sources, the method produces a parameterization of their spatial properties, and can therefore be used in 3D sound synthesis of the recorded mixture.

We evaluate the separation quality of the proposed method against the reference CNMF approach [19] and frequency-domain ICA [8], [29]. The simulations are done using a small microphone array consisting of four microphones enclosed in a casing similar in size to a hand-held mobile device. The evaluation is based on objective separation quality metrics proposed in [30], [31] and perceptually motivated metrics [32]. The proposed method is shown to produce considerably better separation quality over the conventional methods.

The rest of the article is organized as follows. In Section II we present the background of spatial audio processing for sound source separation. The general principle of the proposed SCM model as a superposition of DoA kernels is presented in Section III. Formulation of the proposed SCM model into the CNMF framework and update rules for the model parameter optimization are presented in Section IV. The source reconstruction based on clustering of DoA kernel weights is presented in Section V. The simulations and separation evaluation are given in Section VI. In Section VII we discuss future work for improving the proposed SCM model and 3D sound synthesis using the proposed spatial audio parameterization.

II. Background

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 consist of stating 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

In time domain an array capture consists of a mixture of sound sources convolved with their spatial responses. The mixing model can be described as

$$\tilde{x}_m(t) = \sum_{k=1}^{K} \sum_{\tau} h_{mk}(\tau) s_k(t-\tau) \tag{1}$$

where the mixture $\tilde{x}_m(t)$ consists of $k = 1...K$ sources captured by microphones $m = 1...M$, and the time-domain sample index is denoted by t. The spatial response from source k to microphone m is represented by a mixing filter $h_{mk}(\tau)$ and the single-channel source signals are denoted by $s_k(t)$.

The convolutive mixing model (1) can be approximated in the STFT domain by instantaneous mixing at each frequency bin as

$$\mathbf{x}_{il} \approx \sum_{k=1}^{K} \mathbf{h}_{ik} s_{ilk} = \sum_{k=1}^{K} \mathbf{y}_{ilk} \tag{2}$$

where $\mathbf{x}_{il} = [x_{il1}, \ldots, x_{ilM}]^T$ is the STFT of the capture $\tilde{x}_m(t)$ with analysis window length of $N = 2I-1$, the positive DFT bin frequencies are denoted by $i = 1...I$ and the STFT frame index by $l = 1...L$. The frequency-domain mixing filter is denoted at each frequency bin by $\mathbf{h}_{ik} = [h_{ik1}, \ldots, h_{ikM}]^T$ and the STFTs of the sources are denoted by s_{ilk}. The spatial images of the sources are denoted as $\mathbf{y}_{ilk} = \mathbf{h}_{ik} s_{ilk}$, which are the source signals as seen by the array, i.e. convolved with their spatial impulse responses. The effective length of the mixing filter $h_{mk}(\tau)$ can be several hundreds of milliseconds but its approximation in frequency domain with an analysis window length of tens of milliseconds works well in practice due to the negligible energy after the main reverberant part of the source spatial response.

B. Signal Representation

The proposed method uses SCMs calculated at each time-frequency point as the signal representation. Spatial covariance calculation translates the absolute phase of the mixture to a phase difference between each microphone pair. In the CNMF work by Sawada et. al [19] it was proposed that for the calculation of the SCMs a magnitude square-rooted version of the array capture is used. This ensures that the nonnegative part in the diagonal of the SCM, modeled by the NMF, contains the magnitude spectrum of the mixture and the individual source spectra are approximately additive.

The magnitude square-rooted version $\hat{\mathbf{x}}_{il}$ of the capture $\mathbf{x}_{il} = [x_{il1}, \dots x_{ilM}]^T$ for a time-frequency point (i, l) is obtained as

$$\hat{\mathbf{x}}_{il} \quad = [|x_{il1}|^{1/2} \operatorname{sign}(x_{il1}), \dots, |x_{ilM}|^{1/2} \operatorname{sign}(x_{ilM})]^T \quad (3)$$

where $\operatorname{sign}(z) = z/|z|$ is the signum function for complex numbers. The SCM for a single time-frequency point is obtained from the array capture vector $\hat{\mathbf{x}}_{il} = [\hat{x}_{il1}, \dots, \hat{x}_{ilM}]^T$ as outer product

$$\mathbf{X}_{il} = \hat{\mathbf{x}}_{il}\hat{\mathbf{x}}_{il}^H, \quad (4)$$

where ^{H} stands for Hermitian transpose. Matrices $\mathbf{X}_{il} \in \mathbb{C}^{M \times M}$ for each time frequency point (i, l) point consist of observation magnitude $|\mathbf{x}_{il}| = [|x_{il1}|, \dots, |x_{ilM}|]^T$ in its diagonal $[\mathbf{X}_{il}]_{nn}$, and off-diagonal values $[\mathbf{X}_{il}]_{nm}, n \neq m$ represent the magnitude correlation and phase difference $|x_{iln}x_{ilm}|^{1/2} \operatorname{sign}(x_{iln}x_{ilm}^*)$ between each microphone pair (n, m).

C. Convolutive Mixing Model in Spatial Covariance Domain

The convolutive mixing model defined in Equation (2) can be expressed in the SCM domain by replacing each term by its covariance counterpart. The SCM domain mixing is expressed as

$$\mathbf{X}_{il} \approx \sum_{k=1}^{K} \mathbf{H}_{ik}\hat{s}_{ilk} = \sum_{k=1}^{K} \mathbf{S}_{ilk}, \quad (5)$$

where $\mathbf{H}_{ik}$ is the spatial covariance matrix for each source at each frequency and $\hat{s}_{ilk}$ is the corresponding source magnitude spectrum.

The matrix $\mathbf{H}_{ik} \in \mathbb{C}^{M \times M}$ denotes the source spatial response $\mathbf{h}_{ik}$ expressed in the form of covariance matrix $\mathbf{h}_{ik}\mathbf{h}_{ik}^H$. The complex-valued monoaural source spectrogram s_{ilk} in the SCM domain results to a real-valued power spectrum $s_{ilk}\overline{s_{ilk}}$. Due to the square-rooted STFT used to calculate the observed SCMs, we denote the sources using their magnitude spectra $\hat{s}_{ilk} = (s_{ilk}\overline{s_{ilk}})^{1/2}$. We can approximate the SCMs being additive since the sources are approximately uncorrelated but also sparse, meaning that only a single source is to be active within each time-frequency point [33]. When using the SCM domain representation defined by Equations (3) - (5), the absolute phase of the sources is not significant from the parameter estimation point of view, and we only model the phase differences between all microphone pairs.

Estimating the source magnitudes $\hat{s}_{ilk}$ and the corresponding SCMs denoted by $\mathbf{H}_{ik}$ by turn would provide the desired BSS properties. However estimating $\mathbf{H}_{ik}$ jointly over all frequencies requires a model that ties together the phase difference over frequencies, which is a difficult constraint to be included in the parameter estimation. For the CNMF-based source separation the SCM estimation proposed in [19] relies on the NMF model to enforce magnitudes $\hat{s}_{ilk}$ to correspond to a single source, which is assumed to yield an estimate of $\mathbf{H}_{ik}$ associated to a single source. A direct estimation of the source SCM and variances at each frequency is done for example in [34], but it again requires solving the frequency-wise permutation. The covariance estimation strategy from [34] with NMF as a source magnitude model has been proposed in [22], thus avoiding permutation ambiguity. However, in both cases the spatial properties estimated separately for each STFT frequency bin i do not utilize the fact that the SCM properties are connected by the TDoA of the direct path and early reflections.

III. Proposed Spatial Covariance Matrix Model by Superposition of DoA Kernels

In the case of direct path propagation or anechoic conditions, the source direction with respect to a receiving array corresponds to a specific time delay between the microphones. In beamforming, the TDoA defined by the look direction of a beamformer is used to align the received microphone signals in order to enhance sources originating from the look direction. A single TDoA determines the desired phase difference over frequencies, making the beamformer implemented in the frequency domain able to integrate source evidence over the whole frequency spectrum. Such a concept has not been widely utilized in BSS since it is difficult to include it to the parameter estimation, and the spatial aliasing makes the delays unambiguous at high frequencies.

The proposed DoA-based SCM model can be used to unify the STFT bin dependencies when estimating the source spatial responses, and to avoid optimizing the model parameters individually for each frequency. The difference of the proposed source separation algorithm to beamforming is that the CNMF optimization algorithm is set to fit a collection of predefined DoA kernels (beamforming kernels) to the observed data and that way to find the most likely DoA of the source in question. By defining the SCM model using only direction-dependent parameters, we can utilize the time delay dependency of the spatial covariance values across frequencies in a CNMF algorithm framework for estimating the source magnitude and spatial properties.

A. Time-difference of Arrival

A specific wavefront-arrival direction corresponds to a set of TDoA values between each microphone pair. The TDoA values depend on the geometry of the array and the relationship is shortly explained in the following. We first consider the array illustrated in Figure 1, where one pair of microphones n and m lie on the xy-plane at locations $\mathbf{n}$ and $\mathbf{m}$, respectively. A unit vector $\mathbf{k}_o$ is pointing towards the look direction from the geometrical center $\mathbf{p}$ of the array. For simplicity, we define that the geometrical center of the array is in the origin of the Cartesian coordinate system, i.e. $\mathbf{p} = [0, 0, 0]^T$. The look

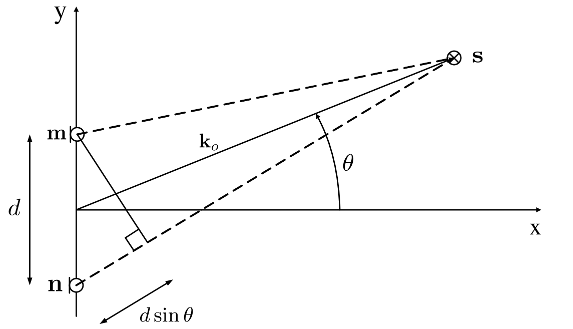

Fig. 1. Example array geometry consisting of two microphones m and n as seen from above, source $\mathbf{s}$ azimuth angle given as θ.

directions can be denoted in the spherical coordinate system using elevation $\theta \in [0, \pi]$, azimuth $\varphi \in [0, 2\pi]$ and fixed radius of $r = 1$. We define ranges of $-90° \leq \theta \leq 90°$ and $0° \leq \varphi \leq 360°$ for elevation and azimuth, respectively.

Assuming the far field model, i.e. the wavefront being planar when arriving to the array, we can write the TDoA of the microphone n with respect to array center point $\mathbf{p}$ in seconds as

$$\tau_n(\mathbf{k}_o) = \frac{-\mathbf{k}_o^T(\mathbf{n} - \mathbf{p})}{v} = \frac{-\mathbf{k}_o^T\mathbf{n}}{v} \quad (6)$$

where v is the speed of sound. Each look direction $o = 1 \ldots O$ translates to a TDoA for each microphone, which further translates into a phase difference linearly proportional to the frequency in the STFT domain. The TDoA in Equation (6) equals to frequency-domain phase difference of $-j2\pi f \tau_n(\mathbf{k}_o)$, where f is the frequency in Hertz. The phase difference is unambiguous only up to the spatial aliasing frequency $f = \frac{v}{2d}$, where d is the smallest distance between any two microphones in the array.

We define TDoA between a microphone pair (n, m) as $\tau_{nm}(\mathbf{k}_o) = \tau_n(\mathbf{k}_o) - \tau_m(\mathbf{k}_o)$. The phase differences corresponding to the TDoA $\tau_{nm}(\mathbf{k}_o)$ between every microphone pair $n = 1 \ldots M$ and $m = 1 \ldots M$ are represented as a matrix $\mathbf{W}_{io} \in \mathbb{C}^{M \times M}$ for each each STFT frequency index $i = 1 \ldots I$ and each look direction $o = 1 \ldots O$. We define these to be DoA kernel matrices which are obtained as

$$[\mathbf{W}_{io}]_{nm} = \exp\big(j2\pi f_i \tau_{nm}(\mathbf{k}_o)\big), \quad f_i = (i-1)F_s/N \quad (7)$$

where the F_s denotes sampling frequency and N is the STFT length.

B. Superposition of DoA Kernels

Assuming a point source and an anechoic capturing condition, a single DoA kernel would be enough to describe the SCM of a source. However, because of echoes and diffractions from surfaces and objects, a more complex model is needed. For SCM modeling, we propose to use a weighted linear combination of DoA kernels that uniformly sample a surface of the unit sphere around the receiving array. The gain of each DoA kernel describe the signal power emanating from each sampled look direction around the array.

We define a set of fixed look directions vectors $\mathbf{k}_o$ that spatially sample the surface of a unit sphere set around the geometrical center $\mathbf{p}$ of the array. An example set of the look direction vectors is illustrated in Figure 2. DoA kernels for each look direction $o = 1 \ldots O$ at each frequency $i = 1 \ldots I$ are denoted using $\mathbf{W}_{io} \in \mathbb{C}^{M \times M}$ and are calculated according to Equation (7). Entries of kernel matrices $[\mathbf{W}_{io}]_{nm}$ denote a TDoA in terms of phase difference expressed as a complex number for a microphone pair (n, m).

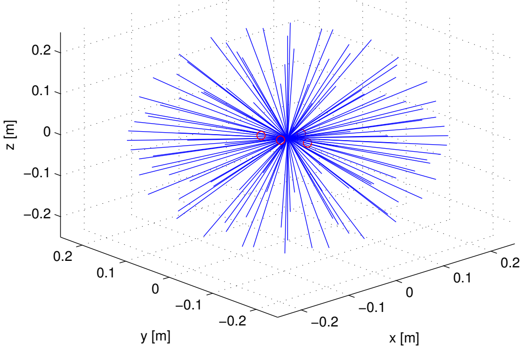

Fig. 2. Look direction vectors approximating uniform sampling of the unit sphere around the array.

In Section II-C the source spatial image was defined as $\mathbf{S}_{ilk} = \mathbf{H}_{ik}\hat{s}_{ilk}$, consisting of the magnitudes $\hat{s}_{ilk}$ and the mixing defined by the source SCM $\mathbf{H}_{ik}$. The proposed SCM model equals the weighted superposition of multiple DoA kernels and is given as

$$\mathbf{H}_{ik} = \sum_{o=1}^{O} \mathbf{W}_{io} z_{ko}, \quad (8)$$

where z_{ko} are the direction weights corresponding to the DoA kernels into each look direction.

We want to estimate $\mathbf{H}_{ik}$ in such a way that it corresponds to a single acoustical source over all the STFT frequencies. This is directly achieved in the proposed SCM model by estimating the spatial weights z_{ko} which are independent of frequency. The definition of the DoA kernels in Equation (7) directly takes into account the frequency dependencies that a certain source DoA causes through a single TDoA. The spatial weights z_{ko} are restricted to be non-negative and they can be estimated in the CNMF framework with a corresponding magnitude model for $\hat{s}_{ilk}$ as will be shown in Section IV. An example of the estimated SCM model weights z_{ko} for three sources are illustrated in Figure 3. The illustration depicts the weighted look direction vectors denoted by $z_{ko}\mathbf{k}_o$ and the result is projected on to the xy-plane. The experimental conditions for obtaining Figure 3 is described in Section VI-A.

IV. Complex-valued Non-negative Matrix Factorization with the Proposed SCM Model

In this section we present a BSS algorithm that combines an NMF-based source magnitude model and the DoA kernel based SCM model which together produce a complex valued NMF (CNMF) model. The proposed BSS algorithm is able to

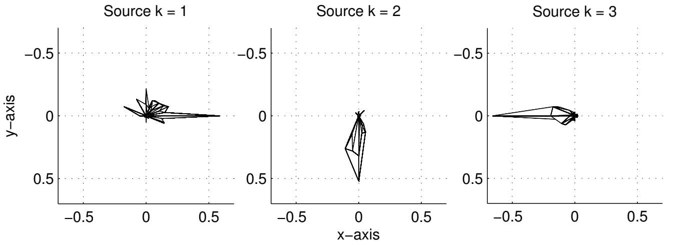

Fig. 3. Illustration of the weighted look direction vectors $z_{ko}\mathbf{k}_o$ of the estimated SCM model projected on to the xy-plane. Sources are at 0, 90 and 180 degrees in azimuth, pictured above the array, and azimuth increasing counterclockwise.

jointly estimate source SCMs across frequencies by using the proposed SCM model and its parameterization of source spatial properties by the direction weights, which are independent of frequency.

The block diagram of the proposed algorithm is given in Figure 4. First the STFT is calculated from the time domain microphone array input and the SCM of each time-frequency point is calculated as defined in Section II-B. The SCM representation serves as an input for the CNMF algorithm. Prior to model parameter estimation a set of DoA kernels with fixed look directions are constructed as defined in Section III-A. The DoA kernels are set to sample the spatial space approximately uniformly around the array. The CNMF algorithm with the proposed DoA-based SCM model is applied to estimate the source parameters, i.e. magnitude spectra and DoA kernel direction weights. In the separation stage the sources are reconstructed from the mixture signal by clustering the components obtained by the CNMF to construct a magnitude mask which is used for obtaining Wiener filter estimates of the source spatial images.

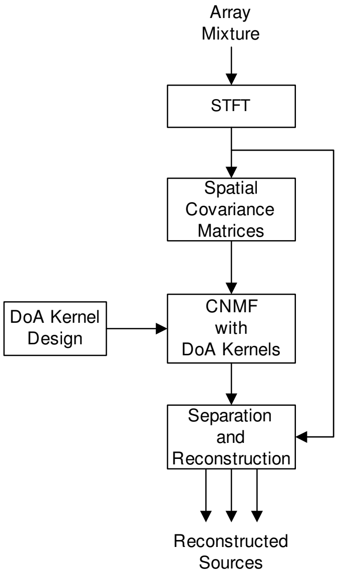

Fig. 4. Block diagram of the proposed BSS system.

A. CNMF Model for SCM Observations

The proposed spatial model consist of a NMF magnitude model [13], [14] for the source magnitude spectra denoted by $\hat{s}_{ilk}$ and the DoA kernel based SCM for denoting the spatial position of the source. In practice, several NMF components are used for representing one actual acoustic sound source, but for the algorithm derivation we define that one NMF component represents one sound source. Later in source signal reconstruction we will cluster the NMF components based on their estimated direction weights z_{ko}.

The model for SCM observations is obtained by replacing $\mathbf{H}_{ik}$ in Equation (5) by the proposed SCM model defined in Equation (8) to get

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{k=1}^{K} \mathbf{H}_{ik}\hat{s}_{ilk} = \sum_{k=1}^{K}\sum_{o=1}^{O} \mathbf{W}_{io} z_{ko} \hat{s}_{ilk}. \quad (9)$$

Source magnitudes $\hat{s}_{ilk}$ are to be obtained by the NMF estimation framework. A rank-1 NMF model for the magnitude spectrogram of a single source k is defined as

$$\hat{s}_{ilk} = t_{ik} v_{kl}, \quad t_{ik}, v_{kl} \geq 0, \quad (10)$$

where column vector $t_{:k}$ contains the spectrum of the source, and the corresponding row $v_{k:}$ represents its gain in each STFT frame. The NMF magnitude model with a fixed source spectrum is extremely simplified and can only model parts of real acoustic sources, but serves as an intermediate representation for the spatial parameter estimation.

Substituting the NMF model (10) into the SCM model (9) and rearranging the parameters gives us the whole CNMF model

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{k=1}^{K}\sum_{o=1}^{O} \mathbf{W}_{io} z_{ko} t_{ik} v_{kl}. \quad (11)$$

Additionally, the CNMF model can be given using the source SCMs $\mathbf{H}_{ik}$ which equals to the model

$$\hat{\mathbf{X}}_{il} = \sum_{k=1}^{K} \mathbf{H}_{ik} t_{ik} v_{kl}. \quad (12)$$

Comparing the models defined in Equations (11) and (12), we observe that the real-valued entries in the diagonal of $\mathbf{H}_{ik}$ are responsible for modeling the absolute source magnitude level with respect to each channel, and the off-diagonal values model the cross-channel magnitude and phase difference properties. This further means that the magnitudes $|\mathbf{W}_{io}|$ combined with the non-negative weights z_{ko} determine the magnitude difference between the channels.

The DoA kernels generated using Equation (7) have unit magnitudes and for modeling the magnitude differences between each channel, the algorithm needs to estimate and update magnitudes of $\mathbf{W}_{io}$ accordingly. This is due to the fact that the sources have gain differences with respect to each microphone. The gain differences are caused by the microphones being at different distance from the source and the possible acoustical shade of the array casing which produces direction-dependent gain even if omnidirectional microphones are used. While the SCM magnitudes are subject to updating, we keep the original DoA kernel phase difference the same, i.e. the original time delay caused by certain direction of the source. In this way we retain the frequency dependency when modeling the phase difference by estimating the frequency independent spatial weights z_{ko}.

B. The CNMF Algorithm

NMF algorithms typically use multiplicative updates iteratively in order to minimize a given cost function, for example the squared Euclidean distance or the Kullback-Leibler divergence [12]. In this paper we present a method for obtaining the algorithm updates via auxiliary functions and EM-algorithm structure similarly as presented in [19].

1) CNMF Cost Function: We aim to minimize the squared Frobenius norm between the observed $\mathbf{X}_{il}$ and the model $\hat{\mathbf{X}}_{il}$ summed over frequency and time indices, which is defined as

$$\sum_{i=1}^{I}\sum_{l=1}^{L} ||\mathbf{X}_{il} - \hat{\mathbf{X}}_{il}||_F^2. \tag{13}$$

In [19] the statistical interpretation of the CNMF model error (13) is shown to be equivalent to the negative log-likelihood (up to terms independent of the model parameters)

$$\mathcal{L}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V}) = \sum_{i=1}^{I}\sum_{l=1}^{L} ||\mathbf{X}_{il} - \sum_{k=1}^{K}\sum_{o=1}^{O} \mathbf{W}_{io} z_{ko} t_{ik} v_{kl}||_F^2. \tag{14}$$

We use this result in deriving the algorithm update rules for optimization of the model parameters $\theta = \{\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V}\}$. We introduce latent components $\mathbf{C}_{ilko}$ given as

$$\mathbf{C}_{ilko} = \mathbf{W}_{io} z_{ko} t_{ik} v_{kl} + r_{ilko}(\mathbf{X}_{il} - \sum_{k,o} \mathbf{W}_{io} z_{ko} t_{ik} v_{kl}), \tag{15}$$

where

$$r_{ilko} = \frac{z_{ko} t_{ik} v_{kl}}{\hat{x}_{il}}, \quad \hat{x}_{il} = \sum_{k,o} z_{ko} t_{ik} v_{kl}. \tag{16}$$

The parameters satisfy $\sum_{k,o} r_{ilko} = 1$ and $r_{ilko} > 0$. The latent components obey

$$\sum_{k=1}^{K}\sum_{o=1}^{O} \mathbf{C}_{ilko} = \mathbf{X}_{il}. \tag{17}$$

Based on techniques introduced in [19] the negative log-likelihood (14) can be minimized using an auxiliary function incorporating the latent components. The auxiliary function is defined as

$$\mathcal{L}^{+}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V},\mathbf{C}) = \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{o=1}^{O} \frac{1}{r_{ilko}} ||\mathbf{C}_{ilko} - \mathbf{W}_{io} z_{ko} t_{ik} v_{kl}||_F^2. \tag{18}$$

According to [19], the likelihood function (18) can be used for an indirect optimization of (14). This is due to the auxiliary function having the properties

$$\mathcal{L}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V}) \leq \mathcal{L}^{+}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V},\mathbf{C}) \tag{19}$$

$$\mathcal{L}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V}) = \min_{\mathbf{C}} \mathcal{L}^{+}(\mathbf{W},\mathbf{Z},\mathbf{T},\mathbf{V},\mathbf{C}), \tag{20}$$

which indicate that minimizing $\mathcal{L}^{+}$ with respect to $\mathbf{W},\mathbf{Z},\mathbf{T}$ and $\mathbf{V}$ corresponds to the minimization of $\mathcal{L}$ which yields optimization of the model parameters with respect to (13). Substituting the definition of $\mathbf{C}_{ilko}$ in Equation (15) to Equation (18) makes it equal to original likelihood (14) and allows indirect optimization of the whole model using the auxiliary variables.

2) Algorithm Updates for the Non-negative Parameters: The derivation of the algorithm updates is achieved via partial derivation of (18) with respect to each model parameter and setting the derivative to zero. The derivations are given in Appendix A. For non-negative model parameters z_{ko}, t_{ik} and v_{kl}, the following update rules are obtained:

$$z_{ko} \leftarrow z_{ko}\Big[1 + \frac{\sum_{i,l} t_{ik} v_{kl} \mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})}{\sum_{i,l} t_{ik} v_{kl} \hat{x}_{il}}\Big] \tag{21}$$

$$t_{ik} \leftarrow t_{ik}\Big[1 + \frac{\sum_{l,o} z_{ko} v_{kl} \mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})}{\sum_{l,o} z_{ko} v_{kl} \hat{x}_{il}}\Big] \tag{22}$$

$$v_{kl} \leftarrow v_{kl}\Big[1 + \frac{\sum_{i,o} z_{ko} t_{ik} \mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})}{\sum_{i,o} z_{ko} t_{ik} \hat{x}_{il}}\Big]. \tag{23}$$

where $\mathbf{E}_{il} = \mathbf{X}_{il} - \sum_{k,o} \mathbf{W}_{io} z_{ko} t_{ik} v_{kl}$ is the error of the model.

3) Algorithm Updates for the SCM Model Parameters: The optimization of the DoA kernels needs a different update scheme, since we desire to retain the phase differences of the predefined kernels but update the relative magnitude differences. For estimation of the DoA kernel magnitudes we first derive the update for complex $\mathbf{W}_{io}$, but restrict the update to its magnitude.

The update rule for $\mathbf{W}_{io}$ via partial derivation is given in Appendix A and results to multiplicative update

$$\hat{\mathbf{W}}_{io} \leftarrow \mathbf{W}_{io}\Big[\sum_{l,k} z_{ko} t_{ik} v_{kl} \hat{x}_{il} + \sum_{l,k} z_{ko} t_{ik} v_{kl} \mathbf{E}_{il}\Big], \tag{24}$$

where $\hat{\mathbf{W}}_{io}$ is a preliminary update with a modified phase difference compared to the actual desired update of magnitudes of $\mathbf{W}_{io}$.

In particular at the highest frequencies the update (24) may produce matrices that are not positive semidefinite. For example, negative values at the diagonal equal to a subtractive magnitude model even though the model assumes purely additive sources. Based on [19] to enforce positive semidefinite matrices an eigenvalue decomposition $\hat{\mathbf{W}}_{io} = \mathbf{V}\mathbf{D}\mathbf{V}^H$ is applied and eigencomponents with negative eigenvalues are set to zero, denoted as $\hat{\mathbf{D}}$. The positive semidefinite matrices are obtained as

$$\hat{\mathbf{W}}_{io} \leftarrow \mathbf{V}\hat{\mathbf{D}}\mathbf{V}^H. \tag{25}$$

For the final update of actual DoA kernels $\mathbf{W}_{io}$ we apply

$$\mathbf{W}_{io} \leftarrow |\hat{\mathbf{W}}_{io}| \exp(i \arg(\mathbf{W}_{io})), \tag{26}$$

which only updates the magnitude part of the DoA kernels and thus the magnitudes of the SCMs.

4) Parameter Scaling: We constrain the scale of DoA kernels as

$$||\mathbf{W}_{io}||_F = 1, \tag{27}$$

which is achieved by applying

$$\mathbf{W}_{io} \leftarrow \frac{\mathbf{W}_{io}}{||\mathbf{W}_{io}||_F} \tag{28}$$

after evaluating the final stage of the update given in Equation (26). The scaling ensures that the SCM part is only responsible of modeling phase differences and relative magnitude differences between the input channels (diagonal values).

Additionally we introduce following constraints for numerical stability:

$$\sum_{o=1}^{O} z_{ko}^2 = 1, \quad \sum_{l=1}^{L} v_{kl}^2 = 1. \tag{29}$$

The scaling of z_{ko} to unity l^2-norm along DoA kernel direction dimension is compensated by multiplying t_{ik} by the same norm. Similarly, enforcing unity l^2-norm to v_{kl} is compensated by scaling of t_{ik}. The scaling of the model parameters is achieved by applying

$$\hat{a}_k = \big(\sum_{l=1}^{L} v_{kl}^2\big)^{1/2}, \quad v_{kl} \leftarrow \frac{v_{kl}}{\hat{a}_k}, \quad t_{ik} \leftarrow t_{ik}\hat{a}_k \tag{30}$$

$$\hat{b}_k = \big(\sum_{o=1}^{O} z_{ko}^2\big)^{1/2}, \quad z_{ko} \leftarrow \frac{z_{ko}}{\hat{b}_k}, \quad t_{ik} \leftarrow t_{ik}\hat{b}_k, \tag{31}$$

after updates of v_{kl} and z_{ko}, respectively.

5) Algorithm Implementation: The proposed CNMF algorithm consists of the following steps.

1) Initialize z_{ko}, t_{ik} and v_{kl} with random values uniformly distributed between zero and one.
2) Initialize $\mathbf{W}_{io}$ according to (7) and apply scaling (28).
3) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
4) Update t_{ik} according to (22).
5) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
6) Update v_{kl} according to (23).
7) Scale v_{kl} to unity l^2-norm and compensate by rescaling t_{ik} as specified in (30).
8) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
9) Update z_{ko} according to (21).
10) Scale z_{ko} to l^2-norm and compensate by rescaling t_{ik} as specified in (31).
11) Recalculate magnitude model $\hat{x}_{il}$ according to (16).
12) Calculate $\hat{\mathbf{W}}_{io}$ according to (24) and enforce it to be positive semidefinite by (25).
13) Update $\mathbf{W}_{io}$ according to (26) and apply scaling (28).

The algorithm is implemented by repeating steps 3-13 for a fixed amount of iterations or until the parameter updates converge.

V. Source Reconstruction and Direction Weights Clustering

The separation of sources corresponding to whole physical entities requires clustering the CNMF components that were earlier interpreted as individual sources. The components span over frequency, but due to their fixed spectrum over time they can only model simple audio objects which need to be clustered based on their spatial orientation. This can be compared to the clustering of ICA components consisting of estimates for single frequency bin, whereas the CNMF components are audio objects that are sematically at a higher level.

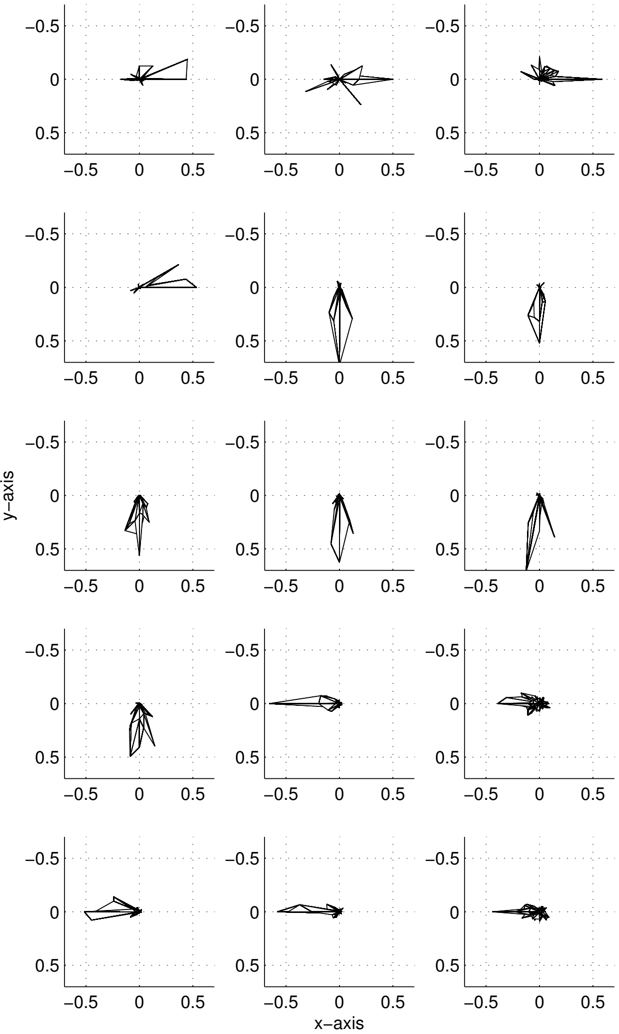

Fig. 5. Look direction vectors weighted by the estimated SCM model parameters for several CNMF components. Vectors denoted by $z_{ko}\mathbf{k}_o$ are projected on to the xy-plane and illustrated as seen above the array. Experimental conditions are described in Section VI-A.

CNMF components originating from the same acoustic source share similar spatial covariance properties determined by their spatial weights z_{ko}. This is illustrated in Figure 5 which depicts SCM model direction weights for several CNMF components showing distinct segmentation to sources at three separate directions. Based on the spatial weight similarity, a separate clustering algorithm can be used to associate CNMF components to the acoustic sources.

We propose to use k-means clustering on the spatial weights z_{ko}. Each $z_{k,:}$ acts as a feature vector and we apply k-means clustering with the cluster count being equal to the number of acoustic sound sources which is defined by the user of the algorithm. We now define the acoustic source index as

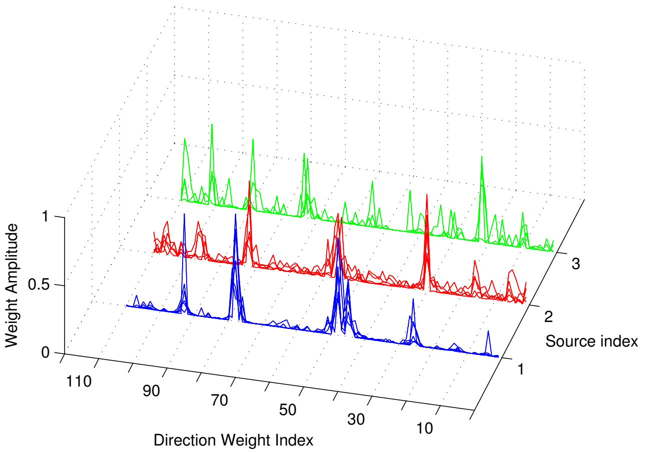

Fig. 6. Direction weights z_{ko} for sources 1, 2 and 3 determined by proposed clustering are illustrated in separated rows. The weights z_{ko} for multiple components k overlaid in each row have a similar peak structure determined by the spatial position of the source.

$q = 1 \ldots Q$. As a result from clustering we get binary cluster decision b_{qk} denoting component k belonging to a source q. An example of direction weights associated to three different sources are illustrated in Figure 6 and direction weights for multiple components k associated to each source are plotted on separate rows. The weights in Figure 6 corresponds to the ones illustrated in Figure 5 projected to the xy-plane.

The CNMF magnitude model for the magnitude spectrogram of an acoustic source q is defined as

$$s_{ilq} = \sum_{ko} b_{qk} z_{ko} t_{ik} v_{kl}. \tag{32}$$

The reconstruction of sources $\mathbf{y}_{ilq}$ as seen by the array, i.e. convolved with their spatial impulse response, based on (32) is given as

$$\mathbf{y}_{ilq} = \mathbf{x}_{il} \frac{\sum_{ko} b_{qk} z_{ko} t_{ik} v_{kl}}{\sum_{qko} b_{qk} z_{ko} t_{ik} v_{kl}}, \tag{33}$$

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

Any other clustering algorithm or CNMF component to source linking strategy can be used to estimate either a binary or a soft decision b_{qk}. We have chosen to use the k-means clustering using spatial weights as features to demonstrate the DoA analysis performance of the proposed SCM model. Other features extracted from the CNMF component parameters such as spectral similarity and gain behavior over time can be used in parallel for associating the CNMF components to the sources to improve the clustering decision [35], [36]. In Section VI-D we study the performance of the chosen k-means clustering strategy against oracle clustering based on known source locations.

VI. Simulations and Source Separation Evaluation

In this section we evaluate the source separation performance of the proposed algorithm. The evaluation consist of a comparison against conventional BSS methods suitable for the case of small microphone array captures with a reasonable amount of reverberation. The separation performance is determined by objective measures, the signal-to-distortion ratio (SDR), image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artefact ratio (SAR). Additionally, perceptually motivated scores proposed in [32] are reported.

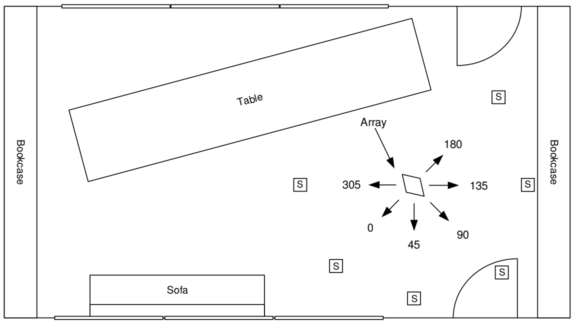

Fig. 7. Capturing room layout, and array and source positions used for datasets one and two.

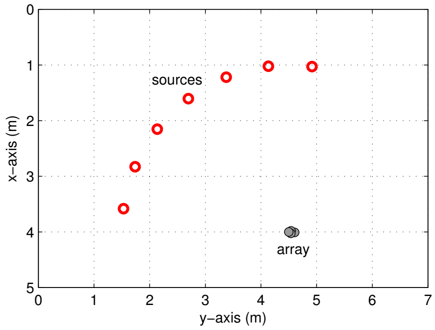

Fig. 8. Simulated room layout for dataset three, microphones illustrated by gray circles and sources by red circles.

A. Evaluation Datasets

For evaluation purposes, a set of room impulse responses (RIR) were measured in a regular meeting room by using an array consisting of four Sennheiser MKE2 omnidirectional condensator microphones inside a metal casing similar to a regular hand-held device. The room dimensions were 7.95 m x 4.90 m x 3.25 m and the reverberation time averaged over all the impulse responses from all locations was $T_{60} = 350$ ms. For obtaining the RIRs, an MLS sequence of order 18 was played using a Genelec 1029 monitor loudspeaker and captured using the array. The room layout and angles of the speaker with respect to the array are given in Figure 7. The angles of the speakers were 0, 45, 90, 135, 180 or 305 degrees, the height of the speaker was set to 1.40 m and the array was placed on a tripod with elevation of 1.08 m. The distance of the loudspeaker to the array was approximately 1.50 m. The microphone locations are given in Table I and the array geometry with a reference axis is illustrated in Figure 9. The spatial aliasing frequency for the given array is 1563 Hz.

Mic	x (mm)	y (mm)	z (mm)
1	0	-46	6
2	-22	-8	6
3	22	-8	6
4	0	61	-18

TABLE I
GEOMETRY OF THE ARRAY USED FOR EVALUATION.

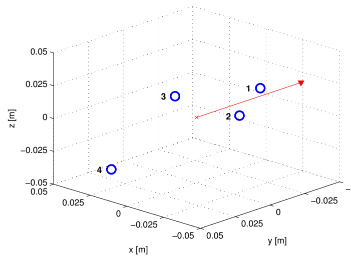

Fig. 9. Array geometry and reference axis.

Case	Source 1	Source 2
1	music 1	music 2
2	male speech 1	power drill
3	male speech 2	hairdryer
4	male speech 3	music 1
5	male speech 1	male speech 2
6	male speech 3	female speech 1
7	female speech 1	movie trailer
8	male speech 2	vacuum cleaner

TABLE II
DESCRIPTION OF SOURCES IN EACH CASE IN DATASET WITH TWO SIMULTANEOUS SOURCES.

Case	Source 1	Source 2	Source 3
1	music 1	music 2	music 3
2	male speech 1	music 1	movie trailer
3	male speech 1	male speech 3	power drill
4	male speech 2	female speech 1	hairdryer
5	male speech 2	male speech 3	music 1
6	male speech 1	male speech 2	female speech 1
7	male speech 2	male speech 3	vacuum cleaner

TABLE III
DESCRIPTION OF SOURCES IN EACH CASE IN DATASET WITH THREE SIMULTANEOUS SOURCES.

The test material was generated by convolving the source-to-array RIRs with varying anechoic source material. The material consist of samples of male and female speech, pop music and various everyday noise sources as listed in Tables II and III. The length of the signals is 10 seconds and original sampling frequency of 48 kHz was downsampled to $F_s = 24$ kHz to reduce the computational complexity. The produced spatial images of sources were summed to obtain an array mixture containing specific sources at specific angles.

Datasets with two and three simultaneous sources were generated, referred as dataset one and two respectively. The datasets consist of cases which refer to combination of different types of sources which are described in Tables II and III. For all cases a set of angles were used which are given in Table IV. The case and angle combinations result to 48 different mixture signals for two simultaneous sources and 42 different mixture signals for three simultaneous sources. It results to eight and seven minutes of evaluation material for datasets one and two, respectively.

Additionally, a simulated room with dimensions of 7 m x 4 m x 3 m was generated using a room simulator based on the source-image method [37]. The simulated room was used to study separation of two sources as the function of the angular spacing between the sources, starting from 15° spacing to 90° with 15° increments. The array was rotated 5° to prevent any special geometry of it being parallel to the walls, and the first source is located at 8° with respect to the reference axis of the array. Distance of sources to the array center is 3 m. The source types used were the same as given in Table II and the corresponding angles are reported in Table IV. The simulated room is illustrated in Figure 8. The target reverberation time was set to 300 ms and the default reflection coefficients for surfaces were used. This test material is referred to as dataset three.

B. Evaluated Methods and Algorithm Parameters

The evaluated algorithms are the proposed CNMF with the DoA kernel based SCM model, the baseline CNMF with unconstrained SCM estimation [19], fullrank SCM estimation [34], ICA variant 1 with magnitude envelope permutation alignment [29] and ICA variant 2 with TDoA permuation alignment [8]. The baseline CNMF was included in the evaluated algorithms to prove the advantages of the proposed DoA kernel based SCM model over the unconstrained SCM model. The ICA methods were chosen to provide comparison to a well known and established BSS techniques.

The results for fullrank SCM estimation [34] are only reported for two sources case. The reference implementation requiring inverting the estimated source covariance matrices which in case of low energy at higher frequencies may become close to singluar and eventually becoming not invertible and preventing the algorithm to proceed from such a state. In the case of three sources dataset, majority of the test samples could not be processed due to the above issue. Additionally, for the same reason one test signal is omitted from the average scores of the fullrank SCM estimation method in two sources dataset.

The proposed separation method only requires three parameters set by the user: the window length, the number of NMF components and the number of iterations for the algorithm updates. The parameters were set to similar values as used in related works [18], [19], and are as follows. The window length of the short-time Fourier transform was set to

	Dataset 1		Dataset 2			Dataset 3	
Angles	1	2	1	2	3	1	2
1	45°	90°	0°	45°	90°	8°	23°
2	135°	180°	45°	90°	135°	8°	38°
3	0°	90°	0°	45°	305°	8°	53°
4	45°	135°	0°	90°	180°	8°	68°
5	0°	135°	0°	135°	180°	8°	83°
6	45°	180°	45°	135°	305°	8°	98°

TABLE IV
ANGLE COMBINATIONS FOR BOTH DATASETS GIVEN IN DEGREES.

$N = 2048$ with 50% window overlap, the window function used was the square root of Hanning window. The number of NMF components for all the CNMF based algorithms was set to $K = 60$ and the algorithms were run for 300 iterations. The true number of sources was given to the methods. ICA and fullrank SCM estimation methods use same STFT analysis parameters given above. For description of the ICA based separation and its method-specific parameters, please refer to [29]. Fullrank SCM estimation was run for 20 iterations.

The look direction vectors for the proposed CNMF are illustrated in Figure 2. It consists of 110 beam directions which sample the unit sphere surface around the array approximately uniformly. The lateral resolution at zero elevation, i.e. at the xy-plane of the array, is 10 degrees, and the different elevations are at 22.5 degrees spacing. The azimuth resolution is decreased close to the poles of the unit sphere.

C. Separation Metrics

The evaluation is done by comparing each separated signal to the spatial images of the original sources and using objective measures by BSS Eval toolbox [30], [31]. Additionally, perceptually motivated scores proposed in [32] are reported.

The discussion of the separation performance is mainly based on signal-to-distortion ratio (SDR) and signal to interference ratio (SIR). The SDR determines how much of the original signal can be explained by the reconstructed source estimates. It is known to emphasize frequency bins with high energy and thus is somewhat dominated by low frequency content especially in case of music samples. The SDR is, however, an established evaluation technique for separation quality comparison. The interference metric SIR determines the amount of cross-talk, and is therefore a good measure of how well each algorithm can separate sources.

Other metrics, SAR and ISR, measure the amount of additional artefacts produced by the separation, and the accuracy of the spatial image of the reconstructed signals, i.e. how well the spatial position of the reconstructed sources is preserved after reconstruction. The used perceptual metrics are overall perceptual score (OPS), target-related perceptual score (TPS), Interference-related Perceptual Score (IPS) and artifact-related perceptual score (APS) [32].

D. Overall Results

The separation scores averaged over all test samples are given in the Tables V and VI for datasets one and two, respectively. The last row labeled as ”mixture” contains the separation metrics evaluated without processing, i.e. calculated for the mixture signal as input for the evaluation toolbox. The results show that the proposed method achieves better average SDR and SIR over the all the compared methods. In the three source dataset the overall separation scores of all the tested methods are fairly low which makes the separation improvement of the proposed method less evident as compared to two sources separation. In the case of three simultaneous sources the SIR performance of the baseline CNMF goes below the performance of ICA separation while the proposed method maintains a better separation with respect to the SIR score. The proposed method also performs best in reconstructing the spatial image of the sources in both test sets.

The score regarding added artefacts to the separated signals measured by the SAR score are lower with the proposed method when compared to the baseline CNMF. This may be attributed to the binary clustering of the proposed method. Faults in the clustering decisions may introduce unwanted rapid changes in the spectrum of the sources, which produces artefacts when reconstructed using the phase of the mixture signal. The baseline CNMF allows soft component-to-source decisions which prevents the added artefacts by smoother spectral discrimination between sources but adds unwanted crosstalk between them. Examples of separated signals from all the evaluated methods are provided at http://www.cs.tut.fi/sgn/arg/nikunen/demo/TASLP2013/.

Method	**SDR** [dB]	**SIR** [dB]	**SAR** [dB]	**ISR** [dB]
CNMF proposed	**4.59**	**7.71**	10.25	**10.29**
CNMF baseline	3.57	4.46	**11.97**	8.31
ICA variant 1	2.86	5.93	9.20	7.87
ICA variant 2	2.03	4.47	8.20	6.95
Fullrank SCM[1]	3.24	6.06	9.19	8.5
Mixture	0.00	0.05	256.76	23.79

Method	**OPS**	**TPS**	**IPS**	**APS**
CNMF proposed	**26.44**	45.70	31.15	54.69
CNMF baseline	16.36	37.13	17.36	**64.01**
ICA variant 1	22.81	46.63	29.99	53.84
ICA variant 2	23.81	46.22	27.83	52.66
Fullrank SCM[1]	25.17	**48.66**	**33.67**	56.86

TABLE V
SEPARATION METRICS FOR DATASET WITH TWO SOURCES.

Method	**SDR** [dB]	**SIR** [dB]	**SAR** [dB]	**ISR** [dB]
CNMF proposed	**2.06**	**4.59**	7.92	**6.37**
CNMF baseline	1.65	-0.10	**9.69**	4.44
ICA variant 1	1.38	3.14	6.39	5.88
ICA variant 2	0.51	1.33	5.59	4.99
Mixture	-3.50	-3.43	251.75	19.62

Method	**OPS**	**TPS**	**IPS**	**APS**
CNMF proposed	21.00	28.51	29.52	40.04
CNMF baseline	17.05	19.64	9.63	**45.10**
ICA variant 1	**26.78**	**49.03**	**37.49**	43.79
ICA variant 2	23.03	44.25	33.95	42.68

TABLE VI
SEPARATION METRICS FOR DATASET WITH THREE SOURCES.

The separation quality measured by the SDR for different cases are reported in Figures 10 and 11 for dataset one and two, respectively. With two simultaneous sources, the proposed method exceeds the baseline CNMF and ICA based separation in most cases. Only in one case the proposed method performs worse than the baseline in terms of the SDR. For very difficult broad-band noise produced by vacuum cleaner, all the tested methods fail to produce adequate separation of the sources. For the three source case, the performance of the evaluated methods have more deviation. The proposed method with simple k-means clustering for determining the NMF component to sources mapping does not produce as constant separation performance as in the case of only two simultaneous

[1] One test signal omitted from the averaging, see Section VI-B.

sources. All the methods fail to provide meaningful separation except in cases 3 and 5.

The overall perceptual score of the proposed method in the case of dataset one indicate the best performance among the tested methods. However, with three simultaneous sources the ICA variant 1 produces better perceptual scores. This can be due to the fact that for most of the cases all the tested methods fail to separate the sources but the reconstruction of sources with ICA based separation are pleasant sounding despite of high amount of crosstalk between the sources. The proposed method in case of faulty component-to-source clustering decision might produce rapid changes in the spectrum which may decrease the perceptual quality and associated scores.

The SDRs for dataset three for the proposed CNMF and baseline CNMF are illustrated in Figure 12 for different angles for the sources. The results clearly indicate that the proposed method benefits from the increased angle between the sources. With the two closest spatial separation of 15 and 30 degrees both the methods produce similar separation with minor advantage for the proposed method. The separation score difference increases when the angle between the sources is increased and starting from 45 degrees the proposed method produces significant improvement for the SDR score over the baseline method.

For performance analysis of the k-means clustering of estimated source spatial weights z_{ko}, a comparison to oracle clustering is provided. The oracle clustering is implemented by searching for the largest value of z_{ko} for each component k and comparing the azimuth of the found index o to the known source angular positions, determined by their azimuth. The component is assigned to the closest known source azimuth. Using the described oracle clustering we only rely on the spatial information of the NMF components but eliminate the effect of possibly faults made by the k-means clustering. The average SDR with oracle clustering are 5.25 dB and 2.94 dB for datasets one and two, respectively. The increase from the k-means clustering are 0.66 dB and 0.88 dB which indicate that the robustness of the k-means clustering is acceptable. Additionally, it can be stated that the spatial weights of the proposed contain the information of the real source spatial positions.

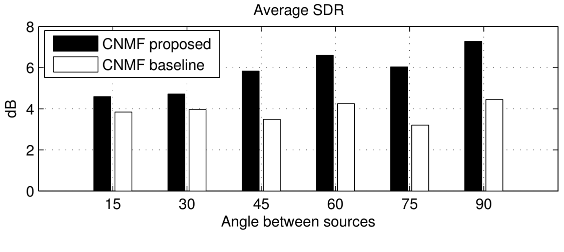

Fig. 12. Averaged SDR for different angle between sources in dataset three with simulated room.

VII. Discussion and Future Work

For the evaluation of the proposed algorithm, a method proposed in [18] that use SCM estimation was also considered. However, it did not prove to be suitable for the given test case according to the separation results and thus no exact separation measures are reported. The lack of any oracle information of the sources and their mixing to give the algorithm a good initial starting point is arguably the reason for low separation performance. Regarding the fullrank SCM estimation [34] of which only partial separation results were reported, issues in solving the permutation ambiguity originating from frequency-wise processing, and the algorithm producing singular SCM estimates for high frequencies were identified as the reason for low separation scores. This indicates the difficulty of the tested case and the efficiency and robustness of the proposed algorithm in analyzing the source spatial covariance.

The future work related to improving the proposed separation framework includes investigating better clustering strategies based on the estimated SCM model direction weights. The k-means clustering does not take into account the geometrical interpretation of the direction weights but solely treats them as feature vectors. Also the clustering decision could be included in the CNMF parameter estimation framework as in done in [19], [21].

In the development of the proposed SCM model, the computational cost of the model was not considered as a design parameter and only good separation performance was sought after. Regarding the computational cost of the model, the number of DoA kernels used for the SCM model increases the computational complexity compared to for example [19]. The average time for performing one iteration with the proposed method takes approximately 9.2 times longer when compared to [19]. The result is obtained with a desktop computer equipped with Intel core2duo E8400 3GHz processor and no special optimization of codes regarding computational complexity for either of the algorithms was made. The computational complexity of the proposed method is approximately linearly proportional to the number of DoA kernels. For example, halving the number of DoA kernels from the used 110 directions to 55 directions decreases the factor to 4.8, when compared to [19]. In general all the CNMF algorithms based on SCM estimation are computationally heavy as they require a high number of iterations in order to converge to a feasible solution.

Reconstruction of the 3D spatial sound field recorded by an irregular array such as the one used in the evaluation requires not only the separation of the sources but information regarding the spatial orientation of the sources. Conventional sound source separation methods do not provide such information, whereas the proposed algorithm can be directly utilized in reconstruction of the spatial sound field the array records. Other soundfield reconstruction methods [38], [39] do not require separation of the sources but utilize more specialized array constellations such as B-format arrays or large linear arrays. For 3D sound field synthesis with the proposed method, the individual CNMF components of the model can be reconstructed without the clustering introduced in Section V, and their direction and spatial spread is determined by the direction weights of the SCM model. Each CNMF component could be synthesized individually and panned and positioned to their analyzed spatial location for example by VBAP [40] or binaural synthesis by HRTF filtering.

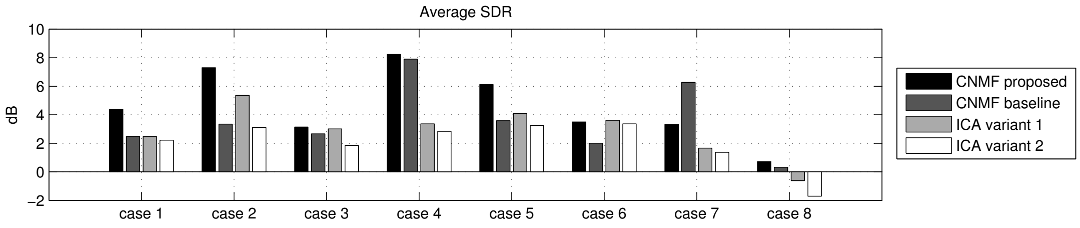

Fig. 10. Averaged SDR for each case for the dataset with two sources.

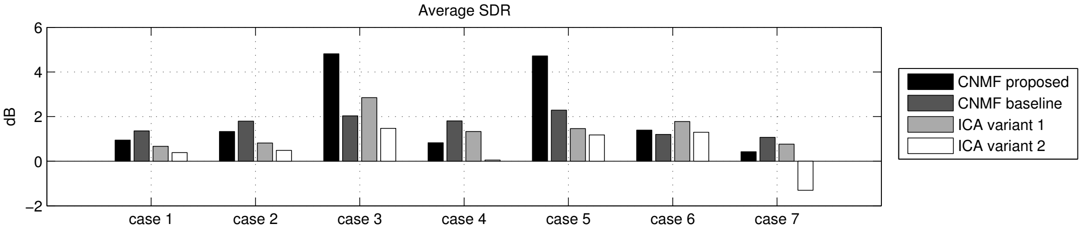

Fig. 11. Averaged SDR for each case for the dataset with three sources.

VIII. Conclusion

In this paper we have proposed a direction of arrival (DoA) based spatial covariance matrix (SCM) model for the purpose of spatial sound source separation using complex-valued non-negative matrix factorization (CNMF). The proposed parameterization of the source SCM by direction-dependent weights allows deriving parameters for the SCM model simultaneously over all frequencies. This improves the overall converge to a spatially coherent solution and mitigates the effect of spatial aliasing which causes problems to many conventional audio separation algorithms. We have shown the separation performance of the proposed algorithm to exceed best performing conventional methods with various types of audio recorded by a small microphone array. The proposed method is a novel approach for spatial parameter estimation in frequency-domain blind source separation, which makes it interesting concept to be utilized in different separation model structures.

Appendix A
Derivation of the CNMF update rules

For the estimation of z_{ko}, t_{ik} and v_{kl}, we redefine the likelihood function (18) by expanding the Frobenius form by using the equality $||\mathbf{A}||^2 = \mathrm{tr}(\mathbf{A}^H\mathbf{A})$ into the form

$$\mathcal{L}^+(\theta) = \sum_{i,l,k,o} \frac{1}{r_{ilko}} [||\mathbf{C}_{ilko}||_F^2 + ||\mathbf{W}_{io}||_F^2 z_{ko}^2 t_{ik}^2 v_{kl}^2 - 2 z_{ko} t_{ik} v_{kl} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})]. \tag{34}$$

Based on the scaling introduced in (28) the second term simplifies to $z_{ko}^2 t_{ik}^2 v_{kl}^2$.

The partial derivatives of (34) with respect to parameters z_{ko}, t_{ik} and v_{kl} are given as

$$\frac{\partial \mathcal{L}^+(\theta)}{\partial z_{ko}} = \sum_{i,l} \frac{2}{r_{ilko}} [z_{ko} t_{ik}^2 v_{kl}^2 - t_{ik} v_{kl} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})] \tag{35}$$

$$\frac{\partial \mathcal{L}^+(\theta)}{\partial t_{ik}} = \sum_{l,o} \frac{2}{r_{ilko}} [z_{ko}^2 t_{ik} v_{kl}^2 - z_{ko} v_{kl} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})] \tag{36}$$

$$\frac{\partial \mathcal{L}^+(\theta)}{\partial v_{kl}} = \sum_{i,o} \frac{2}{r_{ilko}} [z_{ko}^2 t_{ik}^2 v_{kl} - z_{ko} t_{ik} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})], \tag{37}$$

where $\theta = \{\mathbf{W}, \mathbf{Z}, \mathbf{T}, \mathbf{V}, \mathbf{C}\}$. Setting the derivatives to zero, substituting r_{ilko} with its definition in Equation (16), and solving the equations with respect to the parameter to be updated, results to update rules

$$z_{ko} \leftarrow \frac{\sum_{i,l} \hat{x}_{il} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})}{\sum_{i,l} t_{ik} v_{kl} \hat{x}_{il}} \tag{38}$$

$$t_{ik} \leftarrow \frac{\sum_{l,o} \hat{x}_{il} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})}{\sum_{l,o} z_{ko} v_{kl} \hat{x}_{il}} \tag{39}$$

$$v_{kl} \leftarrow \frac{\sum_{i,o} \hat{x}_{il} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io})}{\sum_{i,o} z_{ko} t_{ik} \hat{x}_{il}}. \tag{40}$$

Above updates can be brought into a multiplicative form (Equations (21) - (23)) by substituting the term in the numerators of the above equations as

$$\hat{x}_{il} \mathrm{tr}(\mathbf{C}_{ilko}\mathbf{W}_{io}) = z_{ko} t_{ik} v_{kl} (\hat{x}_{il} + \mathrm{tr}(\mathbf{E}_{il}\mathbf{W}_{io})) \tag{41}$$

and applying some trivial manipulations of the equations.

The update rule for the spatial covariance matrices $\mathbf{W}_{io}$ is obtained via partial derivation of the negative log-likelihood (18) with respect to $\mathbf{W}_{io}$ which is

$$\frac{\partial \mathcal{L}^+(\theta)}{\partial \mathbf{W}_{io}} = \sum_{l,k} \frac{2}{r_{ilko}} (\mathbf{C}_{ilko} - \mathbf{W}_{io} z_{ko} t_{ik} v_{kl})(-z_{ko} t_{ik} v_{kl}). \tag{42}$$

Setting the above derivative to zero and substituting r_{ilko} with its definition in Equation (16) results in the update

$$\hat{\mathbf{W}}_{io} \leftarrow \frac{\sum_{l,k} \hat{x}_{il} \mathbf{C}_{ilko}}{\sum_{l,k} \hat{x}_{il} z_{ko} t_{ik} v_{kl}} \tag{43}$$

Due to the scaling defined in (29) the divisor in the above update can be disregarded. Substituting $\mathbf{C}_{ilko}$ with its definition (15) the above update can be modified into the multiplicative update given in (24).

References

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of Audio Engineering Society*, vol. 55, no. 6, p. 503, 2007.

[2] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and dirac technology," *Journal of Audio Engineering Society*, vol. 59, no. 12, pp. 924–935, 2010.

[3] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2005.

[4] P. Smaragdis, "Extraction of speech from mixture signals," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. John Wiley & Sons, 2012.

[5] J. McDonough and K. Kumatani, "Microphone arrays," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds. John Wiley & Sons, 2012.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. Wiley-interscience, 2001.

[7] M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. 881–884.

[8] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[9] 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.

[10] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2000, pp. 2985–2988.

[11] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[12] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[13] D. Lee, H. Seung *et al.*, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[14] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[15] T. Virtanen, "Monoaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1066–1074, 2007.

[16] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[17] D. FitzGerald, M. Cranitch, and E. Coyle, "Sound Source Separation Using Shifted Non-negative Tensor Factorisation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2006.

[18] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[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. 153–156.

[20] ——, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[21] ——, "Formulations and algorithms for multichannel complex nmf," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 229–232.

[22] 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 *International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[23] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 71–75.

[24] I. Tashev, *Sound capture and processing: practical approaches*. John Wiley & Sons Inc, 2009.

[25] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[26] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[27] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2002.

[28] ——, "Spherical microphone arrays for 3d sound recording," in *Audio Signal Processing: For Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Springer, 2004.

[29] J. Nikunen, T. Virtanen, P. Pertila, and M. Vilermo, "Permutation alignment of frequency-domain ica by the maximization of intra-source envelope correlations," in *20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 1489–1493.

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

[31] 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.

[32] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[33] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[34] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[35] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2005.

[36] T. Barker and T. Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation," in *Proceedings of INTERSPEECH*, 2013.

[37] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, p. 269, 2008.

[38] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Proceedings of 114th Audio Engineering Society Convention*, 2003.

[39] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "Design of transform filter for reproducing arbitrarily shifted sound field using phase-shift of spatio-temporal frequency," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 381–384.

[40] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

Joonas Nikunen (joonas.nikunen@tut.fi) received the M.Sc Degree in signal processing and communications engineering in 2010 from Tampere University of Technology (TUT), Finland. He is currently working as a researcher and post-graduate student at the Department of Signal Processing in TUT and pursuing towards his Ph.D. degree. His research interests include spatial audio representations and coding, sound source separation and object-based representations for audio.

Tuomas Virtanen (tuomas.virtanen@tut.fi) is an Academy Research Fellow and an adjunct professor at Department of Signal Processing, Tampere University of Technology (TUT), Finland. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He has also been working as a research associate at Cambridge University Engineering Department, UK. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition, music content analysis and audio 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 100 scientific publications on the above topics. He has received the IEEE Signal Processing Society 2012 best paper award.