MULTICHANNEL AUDIO SEPARATION BY DIRECTION OF ARRIVAL BASED SPATIAL COVARIANCE MODEL AND NON-NEGATIVE MATRIX FACTORIZATION

Joonas Nikunen, and Tuomas Virtanen

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
joonas.nikunen@tut.fi, tuomas.virtanen@tut.fi

ABSTRACT

This paper studies multichannel audio separation using non-negative matrix factorization (NMF) combined with a new model for spatial covariance matrices (SCM). The proposed model for SCMs is parameterized by source direction of arrival (DoA) and its parameters can be optimized to yield a spatially coherent solution over frequencies thus avoiding permutation ambiguity and spatial aliasing. The model constrains the estimation of SCMs to a set of geometrically possible solutions. Additionally we present a method for using a priori DoA information of the sources extracted blindly from the mixture for the initialization of the parameters of the proposed model. The simulations show that the proposed algorithm exceeds the separation quality of existing spatial separation methods.

Index Terms— Spatial sound separation, non-negative matrix factorization, spatial covariance models

1. INTRODUCTION

Sound source separation has many applications which include for example signal enhancement for speech recognition [1] and object-based audio coding [2]. The separation of multichannel audio is usually based on the estimation of the mixing filter in time or frequency domain. Along with the underlying mixing, there exists spectral structure of the sources that can be analyzed from the mixture for example by non-negative matrix factorization (NMF). The utilization of NMF in separation of spatial audio captures in combination with spatial covariance matrix (SCM) estimation has been studied in [3, 4, 5]. Their benefits over conventional methods such as frequency-domain independent component analysis (ICA) [6, 7, 8] are the absence of permutation problem, and the utilization of audio spectrogram redundancy in estimating audio objects, i.e. NMF components, that span over frequency and time. The previous approaches estimate SCMs separately for each frequency of each source, without placing any constraints on the SCMs.

The unconstrained estimation of source SCMs causes several problems. Estimating SCMs separately for each frequency leads to not only the permutation problem [9], but may also produce solutions that are not spatially coherent. Using the NMF as a source magnitude model introduces frequency dependency, but sources at different spatial locations with similar spectral characteristics may become modeled using a single NMF component. Therefore estimating SCMs for NMF components or a group of them still not guarantee a spatially coherent solution.

In this paper, we introduce a direction of arrival (DoA) based SCM model for spatial audio separation and use NMF as the source magnitude model. We propose to model the source SCMs as a weighted combination of DoA kernels which are derived similarly to array manifold vectors towards a certain look direction as in the field of beamforming [10]. A benefit of the model over ones used in [4, 5, 11] is that the proposed structure of the SCMs is constrained by geometrically possible source directions by knowing the array geometry and phase difference caused by each DoA. Additionally, parameterizing the source SCMs by a set of DoA kernels with fixed look directions and their weights results in a model that unifies the phase difference over frequency and thus its parameters are independent of frequency. The proposed model ensures that the SCM for a source is spatially coherent. Furthermore, conventional DoA analysis tools can be used to initialize its parameters.

This paper proposes an improved version of the system [12] and differs from it by estimating the SCM model for entire sources instead of individual NMF components. Additionally, we propose a blind DoA analysis front-end to initialize the SCM model direction weights. We evaluate the performance of the proposed method compared to the method proposed in [4] and to conventional ICA separation [6].

The rest of the paper is organized as follows. In Section 2 we give the problem definition of spatial source separation and source mixing in the spatial covariance domain. The proposed DoA kernel based SCM model is given in Section 3 and a source DoA estimation front-end for initialization of its parameters is explained in Section 4. A complex-valued NMF model incorporating the direction of arrival based SCM model and the optimization of its parameters is presented in Section 5. Simulations for separation quality evaluation with the proposed method are presented in Section 6.

2. PROBLEM DEFINITION

We assume convolutive mixing of sources in time domain, which is approximated by instantaneous mixing in frequency domain. The mixing model is defined as

$$\mathbf{x}_{il} \approx \sum_{p=1}^{P} \mathbf{h}_{ip} s_{ilp} = \sum_{p=1}^{P} \mathbf{y}_{ilp} \tag{1}$$

where $\mathbf{x}_{il} = [x_{il1}, \ldots, x_{ilM}]^T$ is the short-time Fourier transformed (STFT) mixture signal consisting of M channels, and $i = 1...I$ and $l = 1...L$ are the frequency and frame index, respectively. The source index is denoted by $p = 1...P$ and mixing filters by $\mathbf{h}_{ip} = [h_{ip1}, \ldots, h_{ipM}]^T$. The STFTs of the sources are denoted by s_{ilp}. Sources convolved with their impulse responses are denoted by $\mathbf{y}_{ilp} = \mathbf{h}_{ip} s_{ilp}$.

As proposed in [4] we use magnitude square rooted STFT

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

for the calculation of the spatial covariance matrices $\mathbf{X}_{il} = \hat{\mathbf{x}}_{il}\hat{\mathbf{x}}_{il}^H \in \mathbb{C}^{M\times M}$. With the above definitions the magnitude spectrum of each channel is at the diagonal of $\mathbf{X}_{il}$, and the spatial properties of the mixture are represented by its off-diagonal values, which encode the magnitude cross correlation and the phase difference between each microphone pair. The spatial covariances are invariant of the absolute phase, which allows estimation of their spatial properties by phase difference only.

The mixing model (1) in SCM domain equals to

$$\mathbf{X}_{il} \approx \sum_{p=1}^{P} \mathbf{H}_{ip}\hat{s}_{ilp}, \tag{3}$$

where $\mathbf{H}_{ip}$ is the covariance matrix for each source at each frequency and $\hat{s}_{ilp} = (s_{ilp}\overline{s_{ilp}})^{1/2}$ is the corresponding source magnitude spectrum. The problem now becomes estimating the source spectrum and its covariance matrices in such a way that they correspond to spatially coherent sources.

3. SPATIAL COVARIANCE MATRIX MODEL

The proposed SCM model for a single source consists of a weighted sum of DoA kernels that each correspond to a fixed look direction. Each DoA kernel represent the phase difference of a source at a specific spatial location and is obtained by knowing the array geometry. The DoA kernels sample the spatial space around the array approximately uniformly. By estimating the weights corresponding to each direction, the estimation of SCMs is constrained to a search space of geometrically feasible solutions. Additionally, the direction weights are independent of frequency which further unifies the estimation of phase difference over frequency.

Assuming direct path propagation, a point source at a specific spatial location causes a set of TDoAs between all the microphone pairs, which translates into a phase difference in the frequency domain. We introduce a look direction vector $\mathbf{k}_o$ pointing from the geometric center of the array to the source determined by azimuth φ and elevation θ. By knowing the array geometry, we can calculate the time delays between every microphone pair $n = 1 \dots M$ and $m = 1 \dots M$ a source at this direction causes. This is analogous to finding array steering vectors for a sum-and-delay beamformer.

We denote the time delay between microphone pair (n, m) corresponding to look direction $\mathbf{k}_o$ as

$$\tau_{n.}(\mathbf{k}_o) = (\mathbf{k}_o^T(\mathbf{n} - \mathbf{m}))/v, \tag{4}$$

where v is the speed of sound and $\mathbf{n}$ and $\mathbf{m}$ are vectors representing the locations of microphones n and m, respectively. The time delay translates into a phase difference that is linearly proportional to frequency f_i in Hertz. The spatial covariance matrix of a specific look direction $\mathbf{k}_o$, termed here as the DoA kernel, is given as

$$[\mathbf{W}_{io}]_{nm} = \exp\left(j2\pi f_i \tau_{nm}(\mathbf{k}_o)\right), \quad f_i = (i-1)F_s/N, \tag{5}$$

for each STFT frequency index i. The sampling frequency is denoted by F_s and N is the FFT length.

Each DoA kernel $\mathbf{W}_{io} \in \mathbb{C}^{M\times M}$ has a fixed look direction index by $o = 1 \dots O$ which sample the spatial space around the array approximately uniformly. In case of a point source in anechoic capturing conditions, a single look direction would be enough to describe the SCM of the source using Equation (5). However, due to reverberation and diffraction, a more complex model is needed. We propose to use a weighted superposition of DoA kernels, i.e. point sources, resulting in the proposed SCM model

$$\mathbf{H}_{ip} = \sum_{o=1}^{O} \mathbf{W}_{io} z_{po}, \tag{6}$$

where z_{po} are the direction weights corresponding to DoA kernels into each look direction $\mathbf{k}_o$.

By estimating the direction weights that are independent of frequency, the proposed model produces an estimate of $\mathbf{H}_{ip}$ that is spatially coherent over frequency. We restrict the direction weights z_{po} to be non-negative and in Section 5 we introduce an estimation algorithm for them. An example of the direction weights estimated as described later in Section 5 is illustrated in Figure 1.

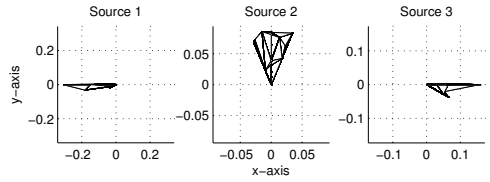

Fig. 1: Illustration of the weighted look direction vectors $z_{po}\mathbf{k}_o$ of the SCM model projected on to the xy-plane. Sources are at 0, 90 and 180 degrees in azimuth.

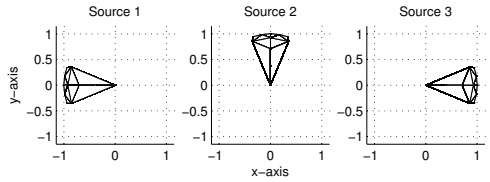

Fig. 2: Illustration of the initialization of the spatial search space for three sources corresponding to Figure 1

4. INITIALIZATION OF DIRECTION WEIGHTS

The parameterization of the source SCM by direction weights z_{po} allows initializing the spatial search space for each source based on DoA analysis of the mixture prior to the model parameter estimation. Based on estimated DoAs defined by azimuth φ_p for each source $p = 1 \dots P$, the direction weights z_{po} are initialized as follows. For each source p the direction weights z_{po} corresponding to look direction indices o within ± 25 degrees from the estimated azimuths φ_p are set to one and all other direction weights of the source are set to zero. The spatial window of 50 degrees accounts for possible errors in the estimation of the source direction in this preprocessing step. An example of the search space used for obtaining the source direction weights in Figure 1 is illustrated in Figure 2.

In the simulations we use the following process to obtain the initial DoA estimates of the sources. Steered response power (SRP) with phase transform (PHAT) [10] is calculated from the STFT of the array signal. The SRP is evaluated for $\varphi = [-180, 180]$ degrees in azimuth with one degree increments and at zero elevation ($\theta = 0$). The maximum of the SRP function at each time frame is scaled to one. The resulting SRP function at each time frame represents the likelihood of a source in each direction. The separation model assumes stationary sources we can therefore average the SRP functions over time. Before averaging, the 15 largest values of the

SRP function are taken from each time frame and the rest of the values are set to zero. Taking only the largest values is equivalent to considering only likelihoods with high confidence. Local maxima that are at least 20 degrees apart from each other are searched from the averaged SRP function. Found locations are set as the initial source DoA estimates. If the number of the found maxima is higher than the number of target sources, the largest maxima are chosen.

5. SEPARATION MODEL

In this section we present the model for the NMF-based spatial sound source separation utilizing the proposed SCM model. Estimation of the parameters of the model follows the framework proposed originally in [4].

The separation model consist of a NMF magnitude model for source spectra $\hat{s}_{ilp} = \sum_{q=1}^{Q} b_{pq} t_{iq} v_{ql}$, where $b_{pq}, t_{iq}, v_{ql} \geq 0$. Each $t_{:q}$ represents the magnitude spectrum of an NMF component, and $v_{q:}$ is its gain in each frame. One NMF component models a single spectrally repetitive event from the mixture and one source is modeled as a sum of multiple components. Parameter b_{pq} represents a soft decision of NMF component q belonging to source p. The second part of the separation model comprises the spatial properties of the sources denoted by $\mathbf{H}_{ip}$, which are represented using the DoA kernel based SCM model $\sum_{o=1}^{O} \mathbf{W}_{io} z_{po}$ as defined in Equation (6). Parameters b_{pq}, t_{iq} and v_{ql} are constrained to non-negative values.

Placing the above definitions into the SCM mixing model defined in Equation (3) results in

$$\mathbf{X}_{il} \approx \hat{\mathbf{X}}_{il} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{o=1}^{O} \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}. \tag{7}$$

The cost function for the parameter optimization is the squared Frobenius norm summed over frequency and time as $\sum_{i,l} ||\mathbf{X}_{il} - \hat{\mathbf{X}}_{il}||_F^2$. As proposed in [4], finding the optimal parameters $\theta = \{\mathbf{W}, \mathbf{Z}, \mathbf{B}, \mathbf{T}, \mathbf{V}\}$ for model (7) is shown to be equivalent to minimizing the following negative log-likelihood

$$\mathcal{L}^{+}(\theta, \mathbf{C}) = \sum_{i,l,p,q,o} \frac{1}{r_{ilpqo}} ||\mathbf{C}_{ilpqo} - \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}||_F^2, \tag{8}$$

with latent components obeying $\sum_{p,q,o} \mathbf{C}_{ilpqo} = \mathbf{X}_{il}$ and being defined as

$$\mathbf{C}_{ilpqo} = \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql} + r_{ilpqo} (\mathbf{X}_{il} - \sum_{q,o} \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}). \tag{9}$$

The parameters $r_{ilpqo} > 0$ are defined as

$$r_{ilpqo} = \frac{z_{po} b_{pq} t_{iq} v_{ql}}{\hat{x}_{il}}, \quad \hat{x}_{il} = \sum_{q,o} z_{po} b_{pq} t_{iq} v_{ql} \tag{10}$$

For optimizing the model parameters multiplicative update equations are derived. The procedure for solving the update rules is based on setting the partial derivatives of (8) with respect to each updated parameter $b_{pq}, z_{po}, t_{iq}, v_{ql}$ and $\mathbf{W}_{io}$ to zero. Substituting $\mathbf{C}_{ilpqo}$ by its definition (9) and applying simple manipulations, this leads to the multiplicative updates

$$b_{pq} \leftarrow b_{pq} \Big[1 + \frac{\sum_{i,l,o} z_{po} t_{iq} v_{ql} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{i,l,o} z_{po} t_{iq} v_{ql} \hat{x}_{il}} \Big] \tag{11}$$

$$z_{po} \leftarrow z_{po} \Big[1 + \frac{\sum_{i,l,q} b_{pq} t_{iq} v_{ql} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{i,l,q} b_{pq} t_{iq} v_{ql} \hat{x}_{il}} \Big] \tag{12}$$

$$t_{iq} \leftarrow t_{iq} \Big[1 + \frac{\sum_{l,p,o} z_{po} b_{pq} v_{ql} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{l,p,o} z_{po} b_{pq} v_{ql} \hat{x}_{il}} \Big] \tag{13}$$

$$v_{ql} \leftarrow v_{ql} \Big[1 + \frac{\sum_{i,p,o} z_{po} b_{pq} t_{iq} \mathrm{tr}(\mathbf{E}_{il} \mathbf{W}_{io})}{\sum_{i,p,o} z_{po} b_{pq} t_{iq} \hat{x}_{il}} \Big], \tag{14}$$

where $\mathbf{E}_{il} = \mathbf{X}_{il} - \sum_{p,q,o} \mathbf{W}_{io} z_{po} b_{pq} t_{iq} v_{ql}$ is the error of the model in each time-frequency point. To ensure numerical stability, the scale of the parameters are normalized as

$$\hat{a}_q = (\sum_{o=1}^{O} z_{po}^2)^{1/2}, \quad z_{po} \leftarrow z_{po} / \hat{a}_q, \quad b_{pq} \leftarrow b_{pq} \hat{a}_q \tag{15}$$

$$\hat{c}_q = (\sum_{l=1}^{L} v_{ql}^2)^{1/2}, \quad v_{ql} \leftarrow v_{ql} / \hat{c}_q, \quad t_{iq} \leftarrow t_{iq} \hat{c}_q. \tag{16}$$

The diagonal entries of $\mathbf{W}_{io}$ model the relative source magnitude level in each channel, and its off-diagonal values model the cross-channel magnitude and phase difference. This means that their unit magnitude as defined by (6) has to be updated in order to model the magnitude level differences in each channel. The update has to maintain the original phase difference, i.e. the original delay caused by a certain look direction.

For updating the magnitudes of $\mathbf{W}_{io}$, we apply the following scheme, also used in [12]. An initial update with a modified phase is calculated as given by the partial derivation of (8)

$$\hat{\mathbf{W}}_{io} \leftarrow \mathbf{W}_{io} \Big[\sum_{l,p,q} b_{pq} z_{po} t_{iq} v_{ql} (\hat{x}_{il} + \mathbf{E}_{il}) \Big]. \tag{17}$$

In order to avoid a subtractive model, matrices $\hat{\mathbf{W}}_{io}$ are forced to be positive semidefinite, which is achieved as proposed in [4] by calculating an eigenvalue decomposition $\hat{\mathbf{W}}_{io} = \mathbf{V}\mathbf{D}\mathbf{V}^H$ and setting negative eigenvalues to zero. Using the modified eigenvalue matrix $\hat{\mathbf{D}}$ the update is reconstructed as $\hat{\mathbf{W}}_{io} \leftarrow \mathbf{V}\hat{\mathbf{D}}\mathbf{V}^H$. The final update preserving the original DoA kernel phase difference is obtained by

$$\mathbf{W}_{io} \leftarrow |\hat{\mathbf{W}}_{io}| \exp(i \arg(\mathbf{W}_{io})). \tag{18}$$

The overall estimation algorithm is implemented as follows. Values of z_{po} are initialized as explained in Section 4 and other parameters are initialized with positive random numbers. The DoA kernels are initialized according to Equation (5). The updates (11) - (14) and (17) - (18) are repeated for a fixed amount of iterations and the parameter scaling as defined by Equations (15) - (16) are applied between iterations. The procedure results in optimizing the model parameters with respect to the squared Frobenius norm between the observations and the model.

The sources are reconstructed as

$$\mathbf{y}_{ilp} = \mathbf{x}_{il} \frac{\sum_{qo} b_{pq} z_{po} t_{iq} v_{ql}}{\sum_{pq'o'} b_{pq'} z_{po'} t_{iq'} v_{q'l}}, \tag{19}$$

which represents Wiener estimates of the sources as seen by the array, i.e. convolved with their spatial impulse responses. The time-domain signals are obtained by inverse STFT and frames are combined by weighted overlap-add.

6. SIMULATIONS

We evaluate the separation quality of the proposed method using separation metrics proposed in [13] and compare its performance against the following methods: NMF with component-wise DoA

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

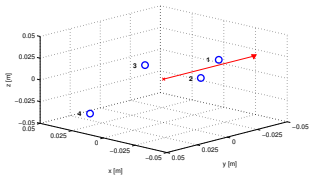

Table 1: Geometry of the array.

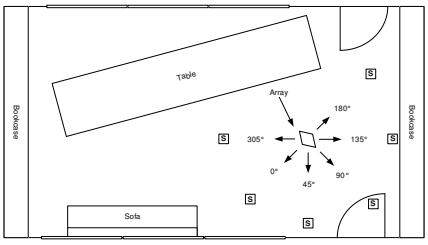

Fig. 3: Room layout, array (tetragon) and source positions (S).

kernel SCM, where the NMF components are grouped to sources by clustering [12], NMF with unconstrained SCM estimation [4] and frequency domain ICA with TDoA based permutation alignment [6].

The test material was generated from anechoic samples that were convolved with room impulse responses (RIR) captured using an array consisting of four omnidirectional microphones enclosed in a metal casing of size 30 mm x 60 mm x 1150 mm. Locations of the microphones are given in Table 1. A Genelec 1029 loudspeaker was used to capture the RIRs from different directions around the array. The height of the loudspeaker 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 recording location was a meeting room with dimensions of 7.95 m x 4.90 m x 3.25 m and the reverberation time averaged over all the impulse responses from all directions was $T_{60} = 350$ ms. The room layout and directions are shown in Figure 3.

The anechoic samples consisted of male and female speech, pop music and various everyday noise sources. The speech samples were obtained from Librivox audiobooks database, the music samples are from RWC Music Genre Database [14] and the noise sources were recorded at an anechoic chamber. Each sample was 10 seconds in duration and they were downsampled from sampling frequency of 48 kHz to $F_s = 24$ kHz. Different datasets for two and three simultaneous sources were generated by convolving the anechoic material with the measured RIRs and summing separate sources from different angles. The used angles are given in Table 2. Using eight combinations of source types for dataset one and seven combinations for dataset two resulted in 48 different mixture signals for two simultaneous sources and 42 different mixture signals for three simultaneous sources.

Dataset 1		**Dataset 2**		
source 1	source 2	source 1	source 2	source 3
45°	90°	0°	45°	90°
135°	180°	45°	90°	135°
0°	90°	0°	45°	305°
45°	135°	0°	90°	180°
0°	135°	0°	135°	180°
45°	180°	45°	135°	305°

Table 2: Angle combinations for both datasets given in degrees.

Method	**SDR**	**SIR**	**SAR**	**ISR**
Proposed	**5.6**	6.8	**13.1**	9.9
NMF clustering [12]	4.8	**8.1**	10.3	**10.5**
NMF Unconstrained [4]	3.7	4.5	12.7	8.4
ICA [6]	2.0	4.5	8.2	6.9

Table 3: Separation metrics for dataset with two sources. All figures in decibels.

Method	**SDR**	**SIR**	**SAR**	**ISR**
Proposed	**3.0**	2.6	**10.7**	6.0
NMF clustering [12]	1.9	**3.8**	7.6	**6.2**
NMF Unconstrained [4]	2.0	0.4	9.9	4.7
ICA [6]	0.5	1.3	5.6	5.0

Table 4: Separation metrics for dataset with three sources. All figures in decibels.

The parameters of the algorithms were set to values similar to the ones used in related studies and are as follows. The window length of the STFT was set to $N = 2048$ with 50% overlap, the window function was square root of Hanning window. The number of NMF components was set to $Q = 60$ and the algorithms were run for 500 iterations. The true number of sources was given to the methods. The DoA kernels for the proposed SCM model consists of 110 directions which sample the unit sphere surface around the array approximately uniformly. The lateral resolution at zero elevation 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.

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). The results averaged over all test samples and all separated sources are given in Tables 3 and 4. The method in [12] is denoted in the tables by "NMF clustering".

The results show that the separation performance of the proposed method exceeds the unconstrained SCM estimation method and frequency-domain ICA across all the measured quantities. The separation measured by SDR when comparing to [4] is increased by 1.9 dB and 1.0 dB in the dataset with two and three sources, respectively. The SIR score denoting source interference is slightly decreased from the NMF component-wise SCM estimation, but it is mostly due to the method in [12] using binary NMF component to source clustering.

7. CONCLUSION

We have presented a spatial audio separation method based on the NMF magnitude model combined with a source SCM model consisting of direction of arrival (DoA) kernels. The strength of the method is the parameterization of the spatial properties of sources by their direction instead of unconstrained estimates which also allows the initialization of the model parameters by a DoA analysis preprocessing step. The separation based on the NMF magnitude model was shown to exceed the quality of the most recent spatial separation method which use unconstrained SCM estimation. An additional benefit of the proposed spatial parameterization is the possibility of the reconstruction of the 3D spatial sound field by positioning the separated sources by their analyzed direction.

8. REFERENCES

[1] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, 2012.

[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] 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.

[4] 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)*. IEEE, 2011, pp. 153–156.

[5] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "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.

[6] 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.

[7] 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*. IEEE, 2008, pp. 43–48.

[8] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation," *Latent Variable Analysis and Signal Separation*, pp. 222–230, 2012.

[9] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Underdetermined 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.

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

[11] S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for underdetermined reverberant audio source separation," in *International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*. IEEE, 2010, pp. 1–4.

[12] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[13] 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.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *Proceedings of the 4th International Conference on Music Information Retrieval*, 2002, pp. 229–230.