PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Source localization using a spatial kernel based covariance model and supervised complex nonnegative matrix factorization

A.J. Muñoz-Montoro[(1)], V. Montiel-Zafra[(1)], J.J. Carabias-Orti[(1)], J. Torre-Cruz[(1)], F.J. Canadas-Quesada[(1)], P. Vera-Candeas[(1)]

[(1)]University of Jaen, Spain

Abstract

This paper presents an algorithm for source localization using a beamforming-inspired spatial covariance model (SCM) and complex non-negative matrix factorization (CNMF). The spatial properties are modeled as the weighted sum of spatial kernels which encode the phase an the amplitude differences between microphones for every possible source location in a grid. The actual localization for each individual source in the multichannel mixture is estimated using complex-valued non-negative matrix factorization (CNMF) where each source spectrogram is modeled using a dictionary of spectral patterns learned a priori from training material.

Localization performance of the proposed system is evaluated using a multi-channel dataset with configurations (number of simultaneous sources, reverberation time, microphones spacing, source types and spatial locations of the sources). Finally, a comparison to other state-of-the-art localization methods is performed, showing competitive localization performance.

Keywords: Localization, DOA estimation, spectral patterns, SCM, CNMF

1 INTRODUCTION

The task of sound source localization (SSL) consists of estimating the spatial positions of the sources in the acoustic scene and has been applied to multiple applications like audio surveillance, teleconferencing, speech enhancement for hearing-aids, or camera pointing systems [14].

Algorithms for SSL can be classified as indirect and direct approaches [11]. Indirect approaches usually follow a two-step procedure: 1) estimate the Time Difference Of Arrival (TDOA) [6] between microphone pairs and, 2) estimate the source position based on the geometry of the array and the estimated delays. Alternatively, direct approaches perform TDOA estimation and source localization in a single step by analyzing a set of candidate locations and selecting the optimal spatial position as an estimate of the source location [7].

In fact, SSL performance is closely related to the quality of the TDOAs estimation. According to [3], there are three general approaches for TDOA estimation: 1) locally for each time-frequency bin and then select the higher peaks in the resulting histogram [8], 2) using an iterative estimation of the time-frequency bins associated to each source and the corresponding TDOAs by clustering method using maximum likelihood [18], 3) estimating an angular spectrum and selecting the high peaks in this function [15]. This angular spectrum can be constructed using different measures, such as correlation between channels [10] or subspace decomposition of the spatial covariance matrix (SCM) [9].

Sound field decomposition aims to represent a sound field as a linear combination of fundamental solutions of the wave equation (or Helmholtz equation) from the results of the pressure measurements. This technique has been applied for several acoustic signal processing applications, such as sound field analysis, reconstruction and visualization. Using this technique, the entire sound field can be estimated from the signals received by multiple microphones which in fact involves source localization or direction of arrival (DOA) estimation because the power distribution of the pressure field indicates the locations or directions of the sound sources.

Several source localization and DOA estimation methods [12, 2, 17] also assume that the soundfield can be represented as a spatially sparse distribution of sound sources over an overcomplete linear equation of the observations. However, in those methods no assumptions are imposed on the structure of the source signals in the time-frequency domain. Alternatively, NMF-based signal decomposition allows further constraints on the

source signal model. For instance, a recent approach in [13] combines supervised CNMF and sparse sound field decomposition, obtaining superior results for DOA estimation than other methods using sparse representations.

In this work we propose a supervised SSL estimation method using a beamforming-inspired SCM model and CNMF. In particular, our method takes the frequency-spectrum structures (spectral patterns) of the source signals into consideration. By exploiting spectral structures trained in advance, it is possible to improve the decomposition accuracy even when the source signals are highly correlated and the sources are in a highly noisy environment. Moreover, this prior information allows to relate each sources and spatial locations without any post-processing stage. The mixing filter for the multichannel input signal is modeled as a weighted combination of the spatial kernels scanning all possible locations in a grid. For the task of DOA estimation, this spatial kernels can be configured to scan all possible directions of arrivals across the 360 degrees [16]. In this work, the spatial kernels are decomposed into two direction dependent SCMs to represent both time and level differences between microphones caused by a single spatial location and its analytic TDOA for a given array geometry [4]. The level differences are represented using a panning-inspired frequency-independent covariance matrix. On the contrary, time delays are modeled using a frequency-dependent phase difference covariance matrix. The source localization is estimated from the frequency-independent directional weights (here denoted as spatial weights) which represent the spatial distribution of the sources across the spatial grid. As a result, the effect of the spatial aliasing is mitigated since the model accounts for phase difference evidence across frequency by single frequency-independent time delays of individual spatial kernels. To estimate the model parameter, namely the activations of the learned spectral patterns for each source, the relative amplitude between microphones and the spatial weights, a CNMF algorithm has been developed using the Itakura Saito optimization function [19].

The estimated spatial location is obtained from the highest peak of the spatial weights associated to each source. Evaluation is performed using the two datasets with different setups in [3] and [5]. Moreover, the proposed method has been compared with other localization methods in the literature in terms of SSL.

The rest of the paper is organized as follows. Section 2 presents the multichannel signal mixing model, section 3 introduces the proposed spatial kernel based SCM localization model. The proposed algorithm for source localization is explained in section 4. Evaluation of the proposed method and comparison with other methods from the literature is performed in section 5. Finally, conclusions are presented in section 6.

2 MULTICHANNEL SIGNAL MIXING MODEL USING SPATIAL COVARIANCE MATRICES

In this work, we use the spatial covariance matrix (SCM) signal representation used in [4, 16, 19]. Rather than absolute phase values, a SCM represents the phase and the amplitude difference between every pair of microphones in the multichannel mixture. Firstly, the magnitude square-rooted matrix $\tilde{x}_{ft}$ for each frequency bin f and time frame t of the captured signal at each sensor $\tilde{x}_{ft} = [\tilde{x}_{ft1}, ..., \tilde{x}_{ftM}]^T$ is computed as

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

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

$$\mathbf{X}_{ft} = \hat{x}_{ft}\hat{x}_{ft}^H = \begin{bmatrix} |\tilde{x}_{ft1}| & \cdots & \tilde{x}_{ft1}\tilde{x}_{ftM}^* \\ \vdots & \ddots & \vdots \\ \tilde{x}_{ftM}\tilde{x}_{ft1}^* & \cdots & |\tilde{x}_{ftM}| \end{bmatrix}, \tag{2}$$

where ^{H} stands for Hermitian transpose. Note that the main diagonal of $\mathbf{X}_{ft}$ encodes the magnitude spectrum of each channel and the spatial properties of the mixture are represented by its off-diagonal values, which encode the phase difference between each microphone pair.

The source mixing model can be defined in terms of the SCM representation as

$$\mathbf{X}_{ft} \approx \hat{\mathbf{X}}_{ft} = \sum_{s=1}^{S} \mathbf{H}_{fs} |\hat{y}_{fts}| \tag{3}$$

where $|\hat{y}_{fts}|$ denotes the magnitude spectrogram for each source $s \in [1,...,S]$ and $\mathbf{H}_{fs} \in \mathbb{C}^{M \times M}$ is the SCM representation of the spatial frequency response $\tilde{\mathbf{h}}_{fs}$. As explained in [16], the SCM model in Eq. 3 can be approximated to be purely additive in anechoic conditions assuming that the sources are uncorrelated and sparse (i.e. only a single source is active at each time frequency (f,t) point). However, this supposition does not always hold in reverberant environments as depicted in [5].

3 PROPOSED SCM MODEL FOR SOURCE LOCALIZATION

In this paper, we present an algorithm for source localization based on SCM and CNMF. In particular, we propose to incorporate within the signal model a dictionary of spectral patterns for all the sources in the mixture. Unlike the method proposed in [5], where the localization is carrying out using the magnitude spectrum for each source, here a previous NMF stage is performed to learn a set of basis for each source and, then, the estimation of the spectrogram parameters in the STFT domain is computed using CNMF.

The proposed signal model for SCM observation is presented in Eq. 4 as

$$\mathbf{X}_{ft} \approx \hat{\mathbf{X}}_{ft} = \sum_{s=1}^{S} \underbrace{\sum_{o=1}^{O} \overbrace{\mathbf{P}_{fo} \circ \mathbf{A}_{o}^{*}}^{\mathbf{W}_{fo}} z_{so}^{*}}_{\mathbf{H}_{fs}} \underbrace{\left(\sum_{k=1}^{K} B_{fks} g_{kts}^{*} \right)}_{|\hat{y}_{fts}|} \tag{4}$$

where the superscript * refers to a free parameter during the factorization and $\circ$ stands for the Hadamard product. The SCM mixing filter $\mathbf{H}_{fs}$ is computed as a linear combination of spatial kernels $\mathbf{W}_{fo} \in \mathbb{C}^{M \times M}$ multiplied by the spatial weights matrix $z_{so} \in \mathbb{R}_{+}^{S \times O}$ which relates sources s with spatial elements of the grid o. Moreover, we propose to decompose the spatial kernels $\mathbf{W}_{fo}$ into two covariance matrices: the phase differences covariance matrix (PDCM) $\mathbf{P}_{fo}$ and the level differences covariance matrix (LDCM) $\mathbf{A}_{o}$.

As in [4], the phase difference between microphones is kept fixed during the factorization process while the relative amplitudes are estimated using the LDCM. In this manner, the parameter $\mathbf{P}_{fo}$ is obtained a priori for every spatial position as $[\mathbf{P}_{fo}]_{pm} = e^{j\theta_{pm}(f,o)}$, where the function $\theta_{pm}(f,o) = 2\pi f_i \tau_{pm}(\mathbf{x}_o)$ encodes the phase difference for each pair of sensors (p,m) at each bin f and spatial position o.

Note that a fixed number of look position $o = 1,...,O$ are used to cover the grid space. For a pair of microphones (p,m), each spatial position $\mathbf{x}_o$ can be turned into a TDOA (in seconds) using the following expression:

$$\tau_{pm}(\mathbf{x}_o) = \frac{||\mathbf{x}_o - \mathbf{p}||_2 - ||\mathbf{x}_o - \mathbf{m}||_2}{c} \tag{5}$$

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

In this work, the magnitude time-frequency spectrogram $|\hat{y}_{fts}|$ for each source s is computed as a linear combination of basis functions $B_{fks} \in \mathbb{R}_{+}^{F \times K \times S}$ and their corresponding time-varying gains g_{kts}. In particular, the basis functions for each source s consist of a set of spectral patterns learned in advance using a standard NMF. During the factorization process, these basis function are fixed.

4 CNMF ALGORITHM FOR SOURCE LOCALIZATION

In this work, the parameters of the proposed SCM model in Eq. 4 are estimated using CNMF. We have used a similar approach to [19] to obtain the multiplicative updates via auxiliary functions for the case of the Itakura Saito (IS) divergence. As demonstrated in [19], MU updates provide faster convergence than the EM algorithms.

4.1 Formulation for Itakura Saito Divergence

The Itakura Saito divergence between the observed $\mathbf{X}_{ft}$ and the estimated $\hat{\mathbf{X}}_{ft}$ SCM observations is expressed as:

$$D_{IS}(\mathbf{X}_{ft},\hat{\mathbf{X}}_{ft}) = tr(\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}) - log(det(\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1})) - M \tag{6}$$

In pursuit of brevity, we write directly the multiplicative update rules of the IS divergence for each free parameter g_{kts} and z_{so} in Eq. 6 as:

$$g_{kts} \leftarrow g_{kts}\sqrt{\frac{\sum_{f,o} z_{so}B_{fks}\ tr(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}{\sum_{f,o} z_{so}B_{fks}\ tr(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}} \qquad z_{so} \leftarrow z_{so}\sqrt{\frac{\sum_{f,t,k} B_{fks}g_{kts}\ tr(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}{\sum_{f,t,k} B_{fks}g_{kts}\ tr(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{W}_{fo})}} \tag{7}$$

Finally, update rules for the level matrix $\mathbf{A}_o$ are obtained as in [19] by solving an algebraic Ricatti equation as $\mathbf{A}_o\mathbf{C}\mathbf{A}_o = \mathbf{D}$, where $\mathbf{C}$ and $\mathbf{D}$ are defined as

$$\mathbf{C} = \sum_{s,k} z_{so}B_{fks}\sum_{f,t} g_{kts}\hat{\mathbf{X}}_{ft}^{-1} \circ \mathbf{P}_{fo}^{*} \qquad \mathbf{D} = \sum_{f}(\mathbf{W}'_{fo}(\sum_{s,k} z_{so}B_{fks}\sum_{t} g_{kts}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1})\mathbf{W}'_{fo}) \circ \mathbf{P}_{fo} \tag{8}$$

Note that $\mathbf{W}'_{fo}$ is the target matrix before the update. The solution of the Riccati equation is obtaining in three step. Firstly, a $2M \times 2M$ matrix $\mathbf{E} = \begin{bmatrix} 0 & -\mathbf{C} \\ -\mathbf{D} & 0 \end{bmatrix}$ is defined. Then, the $2M$ eigenvectors from $\mathbf{E}$ is computed as $e_1,...,e_{2M}$ and, finally, the eigenvectors is sorted according to the associated eigenvalues in ascending order and remove the vectors corresponding to the smallest eigenvalues discarding the negative eigenvalues. As a result, M sorted eigenvectors $e'_1,...,e'_M$ are achieved.

Then, the new $\mathbf{A}_o$ is obtained as $\mathbf{A}_o \leftarrow \mathbf{I}\mathbf{J}^{-1}$, where the $M \times M$ matrices $\mathbf{I}$ and $\mathbf{J}$ are defined from the sorted eigenvectors $\mathbf{e}'$ as $\mathbf{J} = [e'_{1,1:M},...,e'_{M,1:M}]$ and $\mathbf{I} = [e'_{1,M+1:2M},...,e'_{M,M+1:2M}]$. Finally, to compensate for computer arithmetical error, we ensure $\mathbf{A}_o$ is Hermitian by $\mathbf{A}_o \leftarrow \frac{1}{2}(\mathbf{A}_o + \mathbf{A}_o^H)$.

5 EVALUATION

5.1 Experimental Setup

The proposed method is evaluated using two datasets for two-channels and four-channel mixture signals. For the task of DOA estimation, the amount of look directions O conforms an angular grid that covers 180 degrees for the two-channel and 360 degrees for the four-channel dataset, respectively. To alleviate the computational complexity, all the directions have been scanned in the zero elevation plane. In addition, in the four-channel dataset and for the task of source localization, the position of the source is estimated as a specific position into a spatial grid in x, y and z axes, where the grid resolution is set as 50 cm for the three axes.

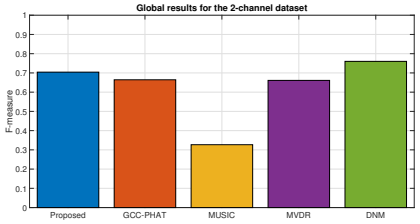
First, a two-channel dataset is composed of 2964 mixtures of 11 s duration corresponding to the male and female speeches belonging to the database used in [3]. The sampling frequency of the mixture signals is 16 kHz. It covers an extensive number of configurations (2 to 6 sources), 4 microphone spacings (5 cm, 15 cm, 30 cm and 1 m), 6 reverberation times RT_{60} (50 ms, 100 ms, 150 ms, 250 ms, 500 ms and 750 ms), 2 source types (male and female speeches) and multiple angular positions of the sources. The mixing filters are generated according to a room of dimensions 4.45 m $\times$ 3.55 m $\times$ 2.5 m.

The four-channel dataset is composed of 96 mixtures of 30 s duration that corresponds to male and female speeches from the original database used in [5]. The sampling frequency of the signals is 16 kHz. This dataset comprises several number of configurations (2 to 3 sources), 2 microphone spacings (5 cm and 54 cm) for the angular grid and 3 microphone positions (located at the ends of the room, and located at the center of the room with spacings of 5 cm and 54 cm) and 2 source types (male and female speeches). The average reverberation time RT_{60} is 350 ms. The room dimensions are 7.95 m $\times$ 4.90 m $\times$ 3.25 m.

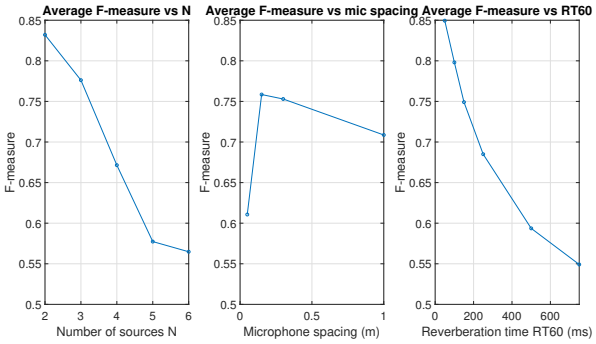
The performance of the proposed algorithm has been compared with five state-of-the-art methods: the generalized cross-correlation with phase transform (GCC-PHAT) [10], the multiple signal classification (MUSIC) [20], the minimum variance distortionless response (MVDR) beamformer [1] and the diffuse noise model (DNM) [3].

Evaluation of the proposed method is performed applying the criteria and evaluation code used in [3]. An estimated DOA is considered as correct if its difference with the true DOA is less than 5 degrees for the angular grid whereas for the spatial grid it is correct if the position is correctly estimated for the three axes. The following metrics are computed: recall R, precision P and F-measure F. These metrics are defined by $R = \frac{I_J}{S}$, $P = \frac{I_J}{J}$ and $F = 2\frac{R \times P}{R+P}$, where J is the number of estimated DOAs, I_J is the number of correct DOAs and finally S is the number of sources. For a fair comparison, the number of sources is assumed to be known ($J = S$). Then, same values for R, P and F are obtained so the latter metric will be evaluated.

5.2 Results for DOA estimation



(a) Average F-measure of the proposed algorithm and the state-of-the art methods for the 2-channel dataset.



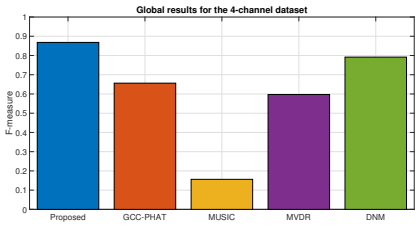
(b) Average F-measure of the proposed algorithm as a function of the number of sources (left), microphone spacing (center) and reverberation time RT_{60} (right) for the 2-channel dataset.

Figure 1. Average results for the 2-channel dataset.

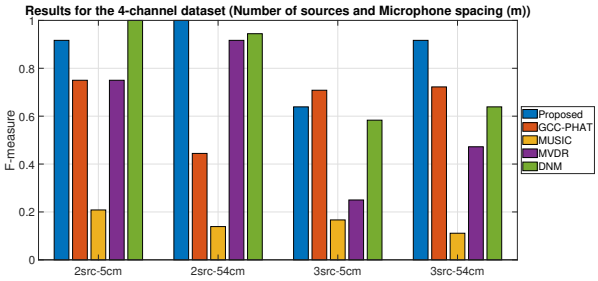
First, the proposed method applying the angular grid is analyzed. Figure 1a shows a comparison with the state-of-the-art methods enumerated in Section 5.1 for the 2-channel dataset. It can be observed that the proposed method and GCC-PHAT, MVDR and DNM methods obtain similar results. For this dataset, DNM method slightly over-performs the proposed method ($F = 0.76$ and $F = 0.7$, respectively).

The F-measure as a function of the number of sources, microphone spacing and reverberation time RT_{60} for the proposed method is presented in Figure 1b. The proposed algorithm obtains higher localization performance when the number of sources decreases. According to the microphone spacing, optimal values are obtained for 15 and 30 cm. Since the proposed algorithm models both the inter-microphone amplitude and the phase difference, when dealing with short spacing arrays (5 cm) the inter-microphone amplitudes are negligible and the localization is performed using the phase information. In the case of large spacing arrays (1 m), the localization is carried out using the amplitude information, because the phase information is ambiguous due to spatial aliasing. Finally, regarding the reverberation time, the localization performance is lower as this time increases.

Figure 2a shows the average F-measure of the proposed and the state-of-the-art methods for the 4-channel



(a) Average F-measure of the proposed algorithm and the state-of-the art methods for the 4-channel dataset.



(b) Average F-measure of the proposed algorithm and the state-of-the art methods for the 4-channel dataset according to the number of sources and the microphone spacing.

Figure 2. Average results for the 4-channel dataset.

dataset. The proposed method obtains the best results among the compared methods ($F = 0.87$). Regarding the other methods, best results are obtained by the DNM method ($F = 0.79$).

A comparison of the proposed method with the state-of-the-art methods according to the number of sources and microphone spacing is presented in Figure 2b. The best localization performance is obtained by the proposed method for the second and fourth setups, where the microphone spacing is 54 cm. For the first and third setups, the proposed method obtains the second best performance.

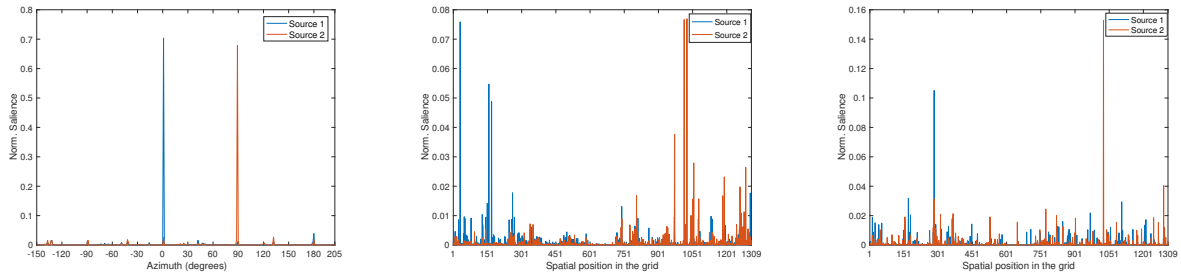
5.3 Results for spatial position estimation in a grid

Here, the proposed method is evaluated for sources located into specific positions into a spatial grid in x, y and z axes, where the grid resolution is set as 50 cm for the three axes. Then, a comparison of the proposed method according to the microphones position is performed. Locating the microphones in the center of the room, the proposed method is not able to correctly estimate the specific position of the source ($F = 0$) for the case of short arrays (5 cm). Dealing with large spacing arrays (1 m), the localization performance improves, obtaining an average of $F = 0.45$ for the cases of 2 and 3 sources. However, when the microphones are placed on the walls of the room, the results outperform, obtaining an average of $F = 0.9$ for the case of 2 sources and $F = 0.4$ for the case of 3 sources. This performance was expected, because, when the array is placed in the center of the room, the distance between the sources and the microphones is large enough to consider a far-field problem, which causes the phase differences between positions with the same DOA to be indistinguishable. However, when the microphones are placed on the walls of the room, a near-field problem is carried out, obtaining better results due to major phase differences for all grid positions.

Figure 3 shows an example of the spatial localization for two concurrent sources. The sources are located into the spatial position in the grid $o = 283$, corresponding to $x = 1$ m, $y = 3$ m, $z = 1$ m and at 0° azimuth; and $o = 1026$, corresponding to $x = 4$ m, $y = 5$ m, $z = 1$ m and at 90° azimuth. For the case of DOA estimation, the proposed method allows to identify the actual DOA for each source. However, it can only identify the correct position in the grid when the microphones are placed on the walls of the room.

6 CONCLUSION

In this work, we proposed a multi-source localization method using a beamforming-inspired SCM model and CNMF with the IS optimization function. The proposed method uses a dictionary of spectral patterns learned in advance for each source, what improves the estimation of the mixing parameters and allows to relate each



(a) Example of DOA estimation for two concurrent sources.

(b) Example of spatial position estimation in a grid for two concurrent sources with the microphone placed in the center of the room.

(c) Example of spatial position estimation in a grid for two concurrent sources with the microphone placed on the walls of the room.

Figure 3. Example of the spatial weights matrix z_{so} estimation for two concurrent sources from the cases: (a) DOA estimation, (b) spatial position estimation in a grid with the microphone placed in the center of the room, and (c) spatial position estimation in a grid with the microphone placed on the walls of the room. The lines in blue and red represents the estimated localization for sources 1 and 2.

sources and spatial positions without any post-processing stage. The mixing model is obtained as a weighted combination of spatial kernels that model the amplitude and the phase differences for every possible source location in a grid.

Localization performance of the proposed system is evaluated using two multichannel datasets with several setups (number of simultaneous sources, reverberation time, microphones spacing, source types and spatial locations of the sources). Finally, a comparison to other state-of-the-art localization methods is performed, showing competitive localization performance.

ACKNOWLEDGEMENTS

This work is supported by Pre-doctoral Fellowship Program from the “Ministerio de Ciencia, Innovación y Universidades” of Spain under the reference BES-2016-078512 and by the University of Jaén under the program “Acción 1. Apoyo a las estructuras de investigación de la Universidad de Jaén para incrementar su competitividad atendiendo a sus singularidades”.

REFERENCES

[1] Araki, S.; Nakatani, T.; Sawada, H.; Makino, S. Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem, Proc. 8th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 2009, pp 742-750.

[2] Asaei A, Bourlard H, Taghizadeh MJ, Cevher V. Model-based sparse component analysis for reverberant speech localization. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2014.

[3] Blandin, C.; Ozerov, A.; Vincent, E.. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering, Signal Processing, Vol 92(8), 2012, pp 1950–1960.

[4] Carabias-Orti, J.J.; Nikunen, J.; Virtanen, T.; Vera-Candeas, P.. Multichannel Blind Sound Source Separation Using Spatial Covariance Model With Level and Time Differences and Nonnegative Matrix Factorization. IEEE/ACM Trans Audio, Speech, Lang Process, September, 2018, Vol 26(9), pp 1512–1527.

[5] Carabias-Orti, J.J.; Cabanas-Molero, P.; Vera-Candeas, P.; Nikunen, J. Multi-source localization using a DOA Kernel based spatial covariance model and complex nonnegative matrix factorization, Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop, Sheffield, UK, 2018, pp 440–444.

[6] Chen J, Benesty J, Huang Y. Time delay estimation in room acoustic environments: An overview. EURASIP J. Appl. Signal Process., vol. 2006, pp. 1–19, 2006.

[7] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Microphone arrays: signal processing techniques and applications," Eds: Michael Brandstein and Darren Ward, Springer-Verlag, 2001.

[8] Faller C, Merimaa J. Source localization in complex listening situations: selection of binaural cues based on interaural coherence. J. Acoust. Soc. Am. 2004

[9] Hassani A, Bertrand A, Moonen M. Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network. Signal Processing. 2015, 107, pp. 68–81

[10] Loesch, B.; Yang, B.. Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions, Proceedings 9th International Conference, LVA/ICA, 2010, pp 41–48.

[11] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," in Advances in Digital Speech Transmission. Hoboken, NJ:Wiley, 2008, pp. 135–166.

[12] Malioutov D, Çetin M, Willsky AS. A sparse signal reconstruction perspective for source localization with sensor arrays. IEEE Trans Signal Process, vol. 53, no. 8, pp. 3010–3022, 2005.

[13] Murata N, Koyama S, Kameoka H, Takamune N, Saruwatari H. Sparse sound field decomposition with multichannel extension of complex NMF. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 395–399, Shanghai, 2016.

[14] Nakano AY, Nakagawa S, Yamamoto K. Automatic estimation of position and orientation of an acoustic source by a microphone array network. J Acoust Soc Am. 126, 3084–3094 (2009).

[15] Nesta F, Svaizer P, Omologo M. Cumulative state coherence transform for a robust two-channel multiple source localization. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009, pp. 290–297.

[16] Nikunen, J; Virtanen, T. Direction of arrival based spatial covariance model for blind sound source separation, IEEE Trans Audio, Speech Lang Process, 2014, Vol 22(3), pp 727–739.

[17] Noohi T, Epain N, Jin CT. Super-resolution acoustic imaging using sparse recovery with spatial priming. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, pp. 2414–2418.

[18] Sawada H, Araki S, Mukai R, Makino S. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. IEEE Trans Audio, Speech Lang Process, vol. 15, no. 5, pp. 1592–1604, July 2007.

[19] Sawada, H.; Kameoka, H.; Araki, S.; Ueda N. Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data, IEEE Trans Audio Speech Lang Processing, 2013, Vol 21(5), pp 971–982.

[20] Schmidt, R. Multiple emitter location and signal parameter estimation, IEEE Transactions on Antennas and Propagation, Vol 34 (3), 1986, pp 276–280.