Interspeech 2022
18-22 September 2022, Incheon, Korea

NTF of Spectral and Spatial Features for Tracking and Separation of Moving Sound Sources in Spherical Harmonic Domain

Mateusz Guzik and Konrad Kowalczyk

Signal Processing Group, Institute of Electronics
AGH University of Science and Technology, Kraków, Poland
mateusz.guzik@agh.edu.pl, konrad.kowalczyk@agh.edu.pl

Abstract

This paper presents a novel Non-negative Tensor Factorization (NTF) based approach to tracking and separation of moving sound sources, formulated in the Spherical Harmonic Domain (SHD). In particular, at first, we redefine an already existing Ambisonic NTF by introducing time-dependence into the Spatial Covariance Matrix (SCM) model. Next, we further extend the time-dependent SCM by incorporating a newly proposed NTF model of the spatial features, thereby introducing spatial components. To exploit the relationship between the positions of sound sources in adjacent time frames, resulting from the naturally occurring continuity of the movement itself, we impose local smoothness on time-dependent components of the spatial features. To this end, we propose a suitable posterior probability with Gibbs prior, and finally we derive the corresponding update rules. The experimental evaluation is based on first-order Ambisonic recordings of speech utterances and musical instruments in several scenarios with moving sources.

Index Terms: source separation, non-negative tensor factorization, spherical harmonic decomposition, spatial Gibbs prior

1. Introduction

The immersive audio and augmented reality applications drive the demand for more advanced sound scene analysis methods, which aim at the estimation of high quality source signals. In context of spatial audio, these methods can be conveniently formulated in the Spherical Harmonic Domain. As a result, much effort has been put towards the advancements in Ambisonic sound source separation, including the NTF [1, 2, 3, 4, 5] and various other approaches [6, 7, 8]. Nevertheless, none of the aforementioned research considers the case of moving sound sources, which is a common scenario in real-life applications. A rare attempt to tackle this issue with regard to Spherical Harmonic Domain was presented in [9], in which the von-Mises modeling of the intensity vector measurements in space and IIR filtering in time were involved.

In this work, we focus on joint sound source tracking and separation problem using Ambisonic NTF. We build our derivations upon the NTF-based SHD SCM model presented in [1], in which the source spectrograms are factorized using the standard NTF, while the individual source SCMs consist of a weighted sum of an additional NTF parameter, referred to as spatial selector, and a set of fixed SH (Spherical Harmonic) Direction-of-arrival (DoA) kernels. First, we reformulate the model to consider a time-dependent SCM, which is done through introduction of the spatial pseudospectrum, by defining the spatial selector weights for each time instant, which yields the TD-SCM. Then we further extend the time-dependent SCM by describing the spatial features with a newly-proposed NTF model, to which we refer as TD-SCM-SC, thereby introducing the idea of spatial components. As a result, we simplify the task of finding appropriate directions in each time frame for all sound sources by dividing the problem into smaller sub-tasks, such that the significant directions, their contribution in time and their association to sound sources are estimated separately. Finally, to utilize the a priori knowledge concerning the local smoothness in time dimension of the spatial pseudospectrum, inferred from the inertial nature of the movement itself, we constrain the newly proposed TD-SCM-SC similarly to [10, 11, 12], thereby deriving the final TD-SCM-SC-GSP method.

We evaluated the proposed solutions with the first-order Ambisonic recordings of speaker utterances and musical instruments. Using the image-source room impulse response simulation technique we created 4 datasets for different movement scenarios and we evaluated the considered methods with respect to their tracking ability, separation quality and convergence. The results of the experimental evaluation clearly indicate that the proposed model along with the smoothness extension significantly enhance both tracking and separation performance.

2. Proposed Ambisonic NTF with spatial components and smoothness constraint

2.1. Mixing model for Ambisonic signals

The Ambisonic mixture of signals emitted by J sound sources in anechoic conditions is given by

$$\mathbf{p}_{ft} = \sum_{j}^{J} S_{jft}\mathbf{y}_j, \tag{1}$$

where $\mathbf{p}_{ft} = [P_{00,ft}, P_{1-1,ft}, P_{10,ft}, P_{11,ft}, \ldots, P_{NN,ft}]^{\mathrm{T}}$ is the L-element vector of the microphone signals in the SHD and S_{jft} is the j-th complex source signal spectrum at frequency f and time frame t. The steering vector $\mathbf{y}_j = [Y_0^0(\Omega_j), Y_1^{-1}(\Omega_j), Y_1^0(\Omega_j), Y_1^1(\Omega_j), \ldots, Y_N^N(\Omega_j)]^{\mathrm{H}}$ associated with j-th DoA is composed of the SH coefficients $Y_n^m(\theta, \phi)$ provided in [13], where n and m denote the SH order and degree, while θ and ϕ are the corresponding colatitude and azimuth angles, respectively.

2.2. NTF with frequency and spatial components

Similarly to [4, 1, 14], the presented NTF framework is oriented towards factorization of the magnitude-compressed spectrogram, which is defined as $\tilde{\mathbf{p}}_{ft} = \left[|P_{1ft}|^{1/2}\sigma(P_{1ft}), \ldots, |P_{Lft}|^{1/2}\sigma(P_{Lft})\right]^{\mathrm{T}}$, where $\sigma(C) = C/|C|$ is the signum function for any complex number C. Apart from ensuring that the values on the diagonal of the SCM consist of the mixture of the approximately additive magnitude spectrum, this approach additionally prevents

10.21437/Interspeech.2022-10526

certain observations from being unnecessarily enhanced, e.g. when applying a scale-sensitive cost function, such as the squared Frobenius norm utilized in this work.

Considering that the reverberant sound field is too complex to be accurately represented by a single anechoic steering vector, in [1] the model of the SHD SCM was defined as a weighted combination of multiple anechoic steering vectors. Nevertheless, our application requires a time-dependent SCM, we therefore rewrite the aforementioned model as

$$\hat{\mathbf{R}}_{ft} = \sum_{j}^{J} \tilde{S}_{jfn} \mathbf{\Xi}_{jt} = \sum_{j}^{J} \tilde{S}_{jfn} \sum_{d}^{D} Z_{jtd} \mathbf{\Sigma}_d, \qquad (2)$$

where $\hat{\mathbf{R}}_{ft}$ models the instantaneous SHD microphone covariance matrix, $\tilde{S}_{jfn} = (S_{jfn} S_{jfn}^*)^{1/2}$ denotes the source magnitude spectrum and $\mathbf{\Xi}_{jt}$ is the time-dependent individual source SCM. The SCMs are expressed through a set of SH DoA kernels $\mathbf{\Sigma}_d$, weighted with $Z_{jtd} \in [0, 1]$, originally referred to as spatial selector and reintroduced here as spatial pseudospectrum. The direction-dependent kernels $\mathbf{\Sigma}_d$ are fixed and defined for a sufficiently high number of directions D distributed uniformly on a spherical manifold.

Analogously to [4, 1], we model the source spectrograms with the NTF [15]

$$\tilde{S}_{jfn} = \sum_{k}^{K} Q_{jk} W_{fk} H_{tk}, \qquad (3)$$

where Q_{jk} assigns the components to sources, W_{fk} contains the frequency profiles and H_{tk} consists of the time activation weights, while K denotes an arbitrary number components.

During preliminary research we established that direct estimation of the spatial pseudospectrum Z_{jtd} is in most cases too overwhelming and results in an improper convergence, therefore we propose to simplify the task, such that the significant directions V_{do}, their contribution in time G_{to} and their association to the sound sources U_{jo} are estimated separately. Thus, we introduce the novel spatial NTF model

$$Z_{jtd} = \sum_{o}^{O} U_{jo} V_{do} G_{to}, \qquad (4)$$

where $o = 1, \ldots, O$ are the spatial components indices.

Altogether, equations (2)-(4) form the proposed model of the instantaneous microphone covariance matrix, that consists of the classical spectral NTF components and the novel spatial NTF components, which is expressed as

$$\hat{\mathbf{R}}_{ft} = \sum_{j}^{J} \sum_{k}^{K} Q_{jk} W_{fk} H_{tk} \sum_{d}^{D} \sum_{o}^{O} U_{jo} V_{do} G_{to} \mathbf{\Sigma}_d \,. \qquad (5)$$

We refer to it as Time-Dependent SCM with Spatial NTF Components (TD-SCM-SC). Through the estimation of the model parameters $\Theta = \{Q_{jk}, W_{fk}, H_{tk}, U_{jo}, V_{do}, G_{to}\}$ the separation of signals emitted by moving sound sources is possible.

2.3. Introduction of the smoothness constraint

Since in most real-life scenarios the assumption concerning close relation of sound source position in consecutive time frames seems reasonable, due to continuous movement rather than very sudden changes of location, we impose local smoothness on the time dimension of the spatial pseudospectrum. This prior knowledge concerning the properties of the sound field can be used to constrain the model by incorporating it in a probabilistic manner, through the Bayesian framework [10, 12].

Therefore, we formulate a posterior probability for the proposed TD-SCM-SC-GSP, with prior on local smoothness introduced in form of the Gibbs distribution as

$$p\left(\Theta|\mathbf{R}_{ft}\right) = \prod \mathcal{N}_c \left([\hat{\mathbf{R}}_{ft}]_{ab} | [\mathbf{R}_{ft}]_{ab}, 1 \right) e^{-\alpha \mathcal{U}(G_{to})}, \quad (6)$$

where $\mathcal{N}_c$ denotes the complex Gaussian distribution, α controls the strength of the Gibbs prior, while $\mathcal{U}(G_{to})$ is a measure of the total roughness of G_{to}, commonly formulated with respect to MRF model in image reconstruction applications. Several suitable choices for $\mathcal{U}$ have been presented in literature [10, 11], of which we choose the Green function [16], and we apply it in the following form

$$\mathcal{U}\left(G_{to}\right) = \frac{\delta}{2\tau} \sum_{t}^{T} \sum_{o}^{O} \sum_{t'} \log\cosh\left(\frac{G_{to} - G_{t'o}}{\delta} \right), \quad (7)$$

where $t' = \{t - \tau, \ldots, t + \tau\} \setminus \{t\}$ is a set of indices that define the local neighbourhood, τ is the number of adjacent time frames and δ is a scaling factor. With this, (7) models the total roughness along the time dimension of G_{to}.

The negative log-likelihood that corresponds to the probabilistic model (6) is given by

$$-\log\left(p\left(\Theta|\mathbf{R}_{ft}\right)\right) = \sum_{f,t}^{F,T} \left\| \hat{\mathbf{R}}_{ft} - \mathbf{R}_{ft} \right\|_{\mathrm{F}}^2 + \alpha \mathcal{U}\left(G_{to}\right), \quad (8)$$

where $\mathbf{R}_{ft} = \mathbf{p}_{ft} \mathbf{p}_{ft}^{\mathrm{H}}$ is the empirical instantaneous covariance matrix and $\|\cdot\|_{\mathrm{F}}^2$ denotes the squared Frobenius norm. Optimization of (8) with respect to the model parameters enables to estimate Θ while considering the smoothness of G_{to}. Note that for $\alpha = 0$, the proposed Time-Dependent SCM with Gibbs Smoothness Prior (GSP) on Spatial NTF Components (TD-SCM-SC-GSP) reduces to TD-SCM-SC and the information concerning smoothness is discarded.

2.4. Derivation of the TD-SCM-SC-GSP update equations

The negative log-posterior (8) can be indirectly optimized using the so-called majorization scheme [14, 17, 18], where a complicated minimization problem is simplified by defining suitable latent components and an appropriate auxiliary function. For our application we define the latent components as

$$\mathbf{C}_{jftkod} = \hat{\mathbf{R}}_{ft}^{-1} Q_{jk} W_{fk} H_{tk} U_{jo} V_{do} G_{to} \mathbf{\Sigma}_d, \qquad (9)$$

with the following property $\sum_{j,k,o,d}^{J,K,O,D} \mathbf{C}_{jftkod} = \mathbf{I}$ and the corresponding auxiliary function

$$\begin{aligned} &\mathcal{L}^{+}\left(\Theta, \mathbf{C}\right) = \\ &\sum_{j,f,t,k,o,d}^{J,F,T,K,O,D} Q_{jk}^2 W_{fk}^2 H_{tk}^2 U_{jo}^2 V_{do}^2 G_{to}^2 \mathrm{tr}\left(\mathbf{\Sigma}_d \mathbf{C}_{jftkod}^{-1} \mathbf{\Sigma}_d\right) - \\ &2 \sum_{j,f,t,k,o,d}^{J,F,T,K,O,D} Q_{jk} W_{fk} H_{tk} U_{jo} V_{do} G_{to} \mathrm{tr}\left(\mathbf{R}_{ft} \mathbf{\Sigma}_d\right) + \\ &\frac{\alpha\delta}{2\tau} \sum_{t}^{T} \sum_{o}^{O} \sum_{t' \in T'(t)} \log\cosh\left(\frac{G_{to} - G_{t'o}}{\delta} \right). \end{aligned} \quad (10)$$

According to [14] it can be shown that minimization of the auxiliary function (10) leads to an indirect minimization of the negative log-posterior (8). The optimization with respect to the model parameters Θ is performed by calculating partial derivatives of (10), which in case of G_{to} is given by

$$\frac{\partial \mathcal{L}^{+}(\Theta, \mathbf{C})}{\partial G_{to}} = \sum_{j,f,k,o}^{J,F,K,O} Q_{jk}^2 W_{fk}^2 H_{tk}^2 U_{jo}^2 V_{do}^2 G_{to} \mathrm{tr}\left(\mathbf{\Sigma}_d \mathbf{C}_{jftkod}^{-1} \mathbf{\Sigma}_d\right) - \sum_{j,f,k,o}^{J,F,K,O} Q_{jk} W_{fk} H_{tk} U_{jo} V_{do} \mathrm{tr}\left(\mathbf{R}_{ft} \mathbf{\Sigma}_d\right) + \frac{\alpha}{4\tau} \sum_{t' \in \mathcal{T}^{(t)}} \tanh\left(\frac{G_{to} - G_{t'o}}{\delta}\right). \quad (11)$$

By transforming equation (11) using the trigonometric identity $\tanh(a-b) = \frac{\tanh(a)-\tanh(b)}{1-\tanh(a)\tanh(b)}$, then substituting $\mathbf{\Sigma}_d \mathbf{C}_{jftkod}^{-1} \mathbf{\Sigma}_d = \left(Q_{jk} W_{fk} H_{tk} U_{jo} V_{do} G_{to}\right)^{-1} \hat{\mathbf{R}}_{ft} \mathbf{\Sigma}_d$ and applying the MU rule [4, 19], the following iterative update equation for G_{to} is obtained

$$G_{to} \leftarrow G_{to} \left[\sum_{j,f,d}^{J,F,D} S_{jft} U_{jo} V_{do} \mathrm{tr}\left(\mathbf{R}_{ft} \mathbf{\Sigma}_d\right) + \frac{\alpha}{2} \xi_{to}\right] \left[\sum_{j,f,d}^{J,F,D} S_{jft} U_{jo} V_{do} \mathrm{tr}\left(\hat{\mathbf{R}}_{ft} \mathbf{\Sigma}_d\right) + \frac{\alpha}{2} \chi_{to}\right]^{-1}, \quad (12)$$

where $\xi_{to} = \frac{1}{2\tau} \sum_t' \left(\tanh^{-1}\left(\frac{G_{t'o}}{\delta}\right) - \tanh\left(\frac{G_{to}}{\delta}\right)\right)^{-1}$ and $\chi_{to} = \frac{1}{2\tau} \sum_t' \left(\tanh^{-1}\left(\frac{G_{to}}{\delta}\right) - \tanh\left(\frac{G_{t'o}}{\delta}\right)\right)^{-1}$.

The update rules for U_{jo} and V_{do} are derived analogously to G_{to} and in their final form they are given by

$$U_{jo} \leftarrow U_{jo} \frac{\sum_{f,t,d}^{F,T,D} S_{jft} V_{do} G_{to} \mathrm{tr}\left(\mathbf{R}_{ft} \mathbf{\Sigma}_d\right)}{\sum_{f,t,d}^{F,T,D} S_{jft} V_{do} G_{to} \mathrm{tr}\left(\hat{\mathbf{R}}_{ft} \mathbf{\Sigma}_d\right)}, \quad (13)$$

$$V_{do} \leftarrow V_{do} \frac{\sum_{j,f,t}^{J,F,T} S_{jft} U_{jo} G_{to} \mathrm{tr}\left(\mathbf{R}_{ft} \mathbf{\Sigma}_d\right)}{\sum_{j,f,t}^{J,F,T} S_{jft} U_{jo} G_{to} \mathrm{tr}\left(\hat{\mathbf{R}}_{ft} \mathbf{\Sigma}_d\right)}, \quad (14)$$

while the update equations for Q_{jk} W_{fk} H_{tk} remain unchanged with respect to those provided in [1].

Similarly to [4, 1], after each update of the parameters associated with the spatial features (12)-(14), a normalization procedure is included. Several normalization schemes were evaluated in a preliminary study, which led us to impose the following set of constraints $\sum_j^J U_{jo} = 1$, $\sum_o^O V_{do} = 1$ and $\sum_o^O G_{to} = 1$. Although these do not assure that $\sum_d^D Z_{jtd} = 1$, this approach provides that the sum remains roughly constant, thus resolving the indeterminacies concerning the confusion of the spatial pseudospectrum and the frequency spectrum.

2.5. Reconstruction of source signals

Given the estimated parameters Θ, we propose to reconstruct the multichannel source images $\hat{S}_{jfn}$ with the Multichannel Wiener Filter (MWF) as

$$\hat{S}_{jfn} = \tilde{S}_{jfn} \mathbf{\Xi}_{jt} \left(\sum_j^J \tilde{S}_{jfn} \mathbf{\Xi}_{jt}\right)^{-1} \mathbf{p}_{ft}. \quad (15)$$

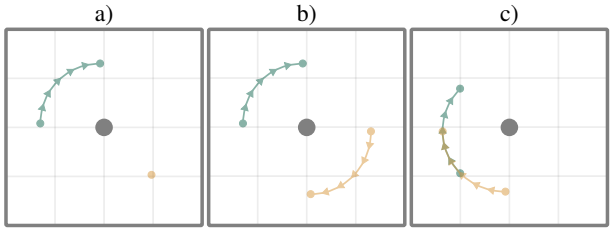

Figure 1: *Experimental setups: subfigures a-c depict source trajectories and microphone array position in scenarios 1-3.*

3. Experimental evaluation

3.1. Experiments and evaluation measures

The experimental evaluation was based on first-order Ambisonic signals, since it continues to be the most popular audio format based on the SHD representation. Nevertheless, the presented derivations remain valid for any SH order and in general as the order increases, the performance improves due to an enhanced spatial selectivity.

The test files were generated using the image-source method [20], such that a microphone array and two sound sources were located inside a 10x8x4 m room with the reverberation time of around 250 ms. A total of four different scenarios were considered, including three with moving sources and one stationary case, while each scenario consisted of 20 5s long files. The three dynamic scenarios are schematically depicted in Fig. 1, in which a) one source is moving on an arc of a circle and one remains stationary, b) both sources are moving (along distinct non-overlapping paths), c) both sources are moving with partial overlap in space. We refer to the static setup as scenario 0, and to three dynamic setups as scenario 1, 2, and 3, respectively.

Two types of source signals were considered, namely speech and musical instruments. The former include utterances of distinct speakers from [21] and the latter consist samples of various types of instruments such as the saxophone, violin, cello, guitar, bass and alto, which were taken from [22].

The proposed TD-SCM-SC-GSP is evaluated against the baseline SCM [1], intermediate TD-SCM and TD-SCM-SC in terms of the localization error, the average signal-to-distortion-ratio (SDR) [23] and the algorithm improvement rate. The localization error was calculated as the Mean Angular Cosine Distance (MACD) between the ground truth and the estimated DoAs at each time instant, with the estimated steering vector given by the normalized first column of the SCM, i.e. as $\hat{\mathbf{y}}_j = \mathbf{\Xi}_{jt}[:,0] / \mathbf{\Xi}_{jt}[0,0]$. Both the SDR and the MACD were averaged over properly converged cases, where the convergence was considered successful if the average SDR obtained with the estimated parameters was improved with respect to the initial parameters. Furthermore, to ensure a straightforward comparison, each algorithm was initialized with the same predefined set of random numbers.

3.2. Results and discussion

Figure 2 presents the ground-truth and the estimated azimuth angles for an example file from scenario 2, with musical instruments as sources. As can be seen, although the TD-SCM performs well with detecting meaningful DoAs, it fails to distinguish them with respect to sound sources. This drawback is lifted in great part with the TD-SCM-SC, while the smoothing

Table 1: *Results of experimental evaluation in scenarios 0-3 for SCM, TD-SCM, TD-SCM-SC and the final proposed TD-SCM-SC-GSP, in case of speech and instrumental signals, in terms of the MACD [°], SDR improvement (ΔSDR) [dB] and improvement rate [%].*

source	scenario	SCM			TD-SCM			TD-SCM-SC			TD-SCM-SC-GSP		
		[°]	[dB]	[%]	[°]	[dB]	[%]	[°]	[dB]	[%]	[°]	[dB]	[%]
speech	0	4.4	9.0	100	42.9	0.3	5	43.8	3.2	75	20.1	8.5	100
	1	14.0	8.0	100	None	None	0	47.6	3.7	65	12.9	9.7	100
	2	25.3	6.3	100	81.3	0.3	15	58.8	2.0	65	6.6	7.1	100
	3	25.2	1.7	90	29.7	0.2	15	27.7	0.9	60	30.6	2.5	65
instruments	0	9.9	11.3	95	37.0	3.6	60	26.1	10.3	100	19.6	11.1	100
	1	14.3	10.0	85	48.3	2.3	45	27.0	8.3	75	16.4	13.7	95
	2	24.9	7.1	85	34.1	2.5	60	22.1	5.9	80	6.3	9.2	95
	3	24.5	1.6	80	27.2	1.0	40	23.0	2.8	45	24.0	2.8	85

constraint within the TD-SCM-SC-GSP helps to further refine the estimated paths.

Table 1 contains the evaluation measures in each experimental setup for all of the considered algorithms. In case of the SCM, the overall performance in dynamic scenarios 1-3 degrades with respect to the static setup of scenario 0, irrespective of the source type. The relatively minor deterioration observed in scenarios 1 and 2 is caused by the SCMs tendency to estimate the source direction roughly in the center of its path. This results in a progressively more accurate estimate as the sound source approaches the middle of the recording and then the opposite tendency toward the end. Therefore, throughout the major part of the audio file a decent localization accuracy is obtained. Consequently, the largest degradation can be seen in scenario 3, in which spatial overlap of source locations occurs. In this case, the fixed look direction combined with the spatial overlap inevitably produces a beampattern that partially captures sound of both sources in the fragments that overlap.

The introduction of time-dependency alone, as is the case for the TD-SCM, in fact negatively affects the separation quality and convergence ability of the method. We hypothesize that this is caused by the aforementioned problem concerning the confusion of sound sources, as presented in Fig. 2 and further reflected in the average localization error. It is our understanding that since the estimated direction needs to be assigned to an individual source in each time frame separately, only a single snapshot can be utilized to compare the spectral content with a broad context.

By the NTF modeling of the spatial features proposed as part of the TD-SCM-SC, the confusion of assigning DoAs to sources is to some degree resolved, partially restoring the convergence ability and reducing the localization error in comparison with the TD-SCM. This effect is most clearly reflected in the results for the instrumental setup.

Extending the TD-SCM-SC with Gibbs smoothness prior on the spatial NTF components leads to a notable increase in separation accuracy, as given by SDR improvement, in dynamic scenarios 1-3. In particular, the smoothness constraint of the proposed TD-SCM-SC-GSP enables to avoid the issue of DoA mismatch among the sources, which is particularly well pronounced in the MACD for the dynamic scenario 2 for both source types. We believe that to obtain a similar magnitude of improvement in each case, the GSP prior strength and its window length would need to be fine-tuned or determined adaptively. Preferably, these parameters should be defined for each source separately, which is possibly a performance limiting factor in case of the scenario 1. We base this supposition both on heuristic experience and on an intuitive guess that different movement types should be characterized by different smoothness setups.

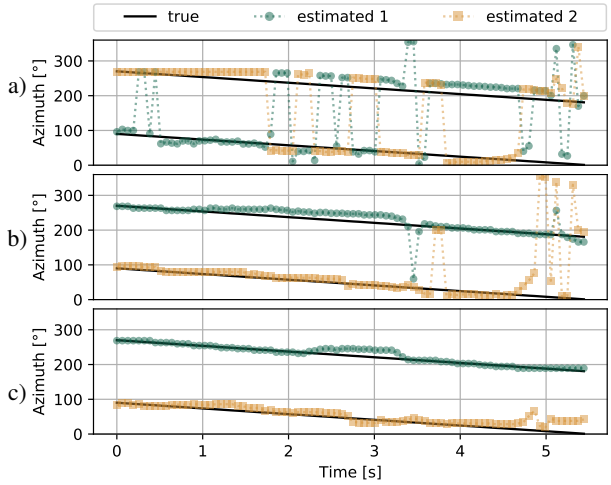

Figure 2: *Example tracking performance in instrumental scenario 2 for: a) TD-SCM, b) TD-SCM-SC, c) TD-SCM-SC-GSP.*

4. Conclusions

In this work, we have proposed an NTF-based method for tracking and separation of moving sound sources. After introducing time dependence into the state-of-the-art Spatial Covariance Matrix model formulated in the Spherical Harmonic Domain, we proposed a novel NTF model for the spatial features, which enables a more convenient estimation of the separate components of the spatial pseudospectrum. The final proposed method, which additionally incorporates local smoothness along the time dimension of the spatial pseudospectrum, significantly improves the localization and tracking capability, which in turn results in better separation efficacy, in dynamic scenarios with moving sound sources. The gain offered by the proposed approach with respect to state-of-the-art is confirmed by the results of experiments performed using Ambisonic recordings of speech and musical instruments.

5. Acknowledgements

We thank Mieszko Fraś and Szymon Woźniak for their helpful remarks. This research was supported by the National Science Centre under grant number DEC-2017/25/B/ST7/01792.

6. References

[1] J. Nikunen and A. Politis, "Multichannel nmf for source separation with ambisonic signals," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 251–255.

[2] M. Hafsati, N. Epain, R. Gribonval, and N. Bertin, "Sound source separation in the higher order ambisonics domain," in *DAFx 2019-22nd International Conference on Digital Audio Effects*, 2019, pp. 1–7.

[3] Y. Mitsufuji, N. Takamune, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[4] M. Guzik and K. Kowalczyk, "Wishart localization prior on spatial covariance matrix in ambisonic source separation using non-negative tensor factorization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[5] M. Guzik, M. Fraś, and K. Kowalczyk, "Incorporation of localization information for sound source separation in spherical harmonic domain," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2021, pp. 1–6.

[6] V. Varanasi and R. Hegde, "Stochastic online dictionary learning for speech source localization and separation in spherical harmonic domain," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 66–70.

[7] S. N. Kalkur, S. Reddy C, and R. M. Hegde, "Joint source localization and separation in spherical harmonic domain using a sparsity based method," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] N. Epain and C. T. Jin, "Independent component analysis using spherical microphone arrays," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 91–102, 2012.

[9] A. Riaz, X. Shi, and A. Kondoz, "Adaptive blind moving source separation based on intensity vector statistics," *Speech Communication*, vol. 113, pp. 1–14, 2019.

[10] R. Zdunek, "Improved convolutive and under-determined blind audio source separation with mrf smoothing," *Cognitive Computation*, vol. 5, no. 4, pp. 493–503, 2013.

[11] R. Zdunek and T. M. Rutkowski, "Nonnegative tensor factorization with smoothness constraints," in *International Conference on Intelligent Computing*. Springer, 2008, pp. 300–307.

[12] R. Zdunek and A. Cichocki, "Gibbs regularized nonnegative matrix factorization for blind separation of locally smooth signals," in *15th IEEE International Workshop on Nonlinear Dynamics of Electronic Systems (NDES 2007), Tokushima, Japan*, 2007, pp. 317–320.

[13] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*. Springer, 2007, vol. 348.

[14] 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.

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

[16] P. J. Green, "Bayesian reconstructions from emission tomography data using a modified em algorithm," *IEEE transactions on medical imaging*, vol. 9, no. 1, pp. 84–93, 1990.

[17] J. De Leeuw, "Block-relaxation algorithms in statistics," in *Information systems and data analysis*. Springer, 1994, pp. 308–324.

[18] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: theory of majorization and its applications*. Springer, 1979, vol. 143.

[19] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2001.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[21] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[22] D. Thery and B. Katz, "Anechoic audio and 3d-video content database of small ensemble performances for virtual concerts," in *Intl Cong on Acoustics (ICA)*, 2019.

[23] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.