WISHART LOCALIZATION PRIOR ON SPATIAL COVARIANCE MATRIX IN AMBISONIC SOURCE SEPARATION USING NON-NEGATIVE TENSOR FACTORIZATION

Mateusz Guzik and Konrad Kowalczyk

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

ABSTRACT

This paper presents an extension of the existing Non-negative Tensor Factorization (NTF) based method for sound source separation under reverberant conditions, formulated for Ambisonic microphone mixture signals. In particular, we address the problem of optimal exploitation of the prior knowledge concerning the source localization, through the formulation of a suitable Maximum a Posteriori (MAP) framework. Within the presented approach, the magnitude spectrograms are modelled by the NTF and the individual source Spatial Covariance Matrices (SCM) are approximated as a sum of anechoic Spherical Harmonic (SH) components, weighted with the so-called spatial selector. We constrain the SCM using the Wishart distribution, which leads to a new posterior probability and in turn to the derivation of the extended update rules. The proposed solution avoids the issues encountered in the original method, related to the empirical binary initialization strategy for the spatial selector weights, which due to multiplicative update rules may result in sound coming from certain directions not being taken into account. The proposed method is evaluated against the original algorithm and another recently proposed Expectation Maximization (EM) algorithm that also incorporates a spatial localization prior, showing improved separation performance in experiments with first-order Ambisonic recordings of musical instruments and speech utterances.

Index Terms— source separation, non-negative tensor factorization, ambisonics, spherical harmonics, array signal processing

1. INTRODUCTION

The rapidly developing immersive audio and augmented reality applications drive the ever increasing demand for more advanced sound scene analysis methods, which enable the estimation of high quality source signals. In context of spatial audio, these methods can be conveniently formulated in the Spherical Harmonic Domain (SHD). In this work, we focus on the Ambisonic sound source separation problem using Non-negative Tensor Factorization.

Very recently, two state-of-the-art NTF-based EM algorithms for sound source separation have been reformulated to operate in the SHD [1]. Moreover, within them the a priori information about the Direction of Arrival (DoA) associated with the localization of the sound source was incorporated using the Gaussian Localization Prior (GLP). In other work, a solution that is suitable for both near and far-field source separation was derived by excluding the noisy evanescent regions from the SHD NTF cost function [2]. In [3] the authors employed the Flexible Audio Source Separation Toolbox (FASST) [4] to investigate the application of the local Gaussian model in case of the source separation in the SHD. Another approach, presented in [5], combines the NTF and the spatial information through the SHD SCM model. This model is based on the approximation of the individual source SCM as a weighted sum of SH components, which results in an additional NTF parameter, referred to as spatial selector. Furthermore, the sparsity based method with an orthonormal constraint on the sparsity matrix was proposed in [6], for joint localization and sound source separation in the SHD. In [7] the stochastic online dictionary learning for speech source localization and separation in the SHD has been proposed. Finally, [8] employs Independent Component Analysis (ICA) for the sound source separation in the SHD.

In this paper, we extend the method that exploits the SHD SCM model presented in [5], which we refer to as SH-SCM-NTF, and we derive a suitable Maximum a Posteriori framework for the incorporation of prior spatial information. The proposed solution is motivated by the fact that the original SH-SCM-NTF method is based on empirical initialization strategy for the spatial selector, in which the a priori knowledge of DoAs corresponding to source locations is used to set the spatial selector weights for directions laying in the vicinity of these DoAs to 1, while setting all other spatial selector weights to 0, such that there is no spatial overlap. There are two major drawbacks of this approach. Firstly, an angular distance threshold must be heuristically determined. Secondly, since Multiplicative Update (MU) rule is used with regard to spatial selector weights, the initially zero-weighted directions remain unused for the entire processing, possibly preventing more accurate estimation of the spatial properties of the sound field, or even diminishing it if the true DoA falls outside the area specified by the threshold, such as might be the case for moving sources. To lift these limitations, we focus on the optimal exploitation of the a priori DoA information by deriving a suitable MAP estimator. By constraining the individual source SCMs with the Wishart Localization Prior (WLP), we formulate the extended optimization problem and subsequently derive new MU rules for the proposed SH-SCM-NTF-WLP algorithm. In conducted experiments, the performance of the proposed algorithm is compared against the original SH-SCM-NTF and the Spherical Harmonic Generalized Expectation Maximization with Gaussian Localization Prior (SH-GEM-GLP) [1], which is a recent state-of-the-art sound source separation technique that also incorporates the information about the source localization through a localization prior. The formulation of the posterior probability with the spatial localization prior and derivation of the proposed method are presented in Sec. 2. The experimental evaluation and conclusions are provided in

This research was supported by the National Science Centre under grant number DEC-2017/25/B/ST7/01792.

Secs. 3 and 4, respectively.

2. PROPOSED AMBISONIC NTF WITH WISHART PRIOR ON SPATIAL COVARIANCE MATRIX

2.1. Mixing model for Ambisonic signals

The acoustic pressure captured with a spherical microphone array can be conveniently represented in the Spherical Harmonic Domain [9], such that the Ambisonic mixture of signals emitted by J sound sources in anechoic conditions is given by

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

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_{jfn} 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 the j-th DoA $\Omega_j = (\theta_j, \phi_j)$ is composed of the SH coefficients, which are expressed as

$$Y_n^m(\Omega) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}\mathcal{P}_n^m(\cos(\theta))\, e^{\jmath m\phi}, \tag{2}$$

where $\mathcal{P}_n^m$ is the associated Legendre polynomial, $\jmath$ is the imaginary unit, while θ and ϕ are the corresponding colatitude and azimuth angles, respectively. This simple SHD plane-wave model is a foundation for further derivation addressing the problem of sound source separation in reverberant environment, as presented in the next section.

2.2. Ambisonic NTF with Spatial Covariance Matrix model

Since the estimation of complex-valued source spectra is in general problematic, in this work the quantity subject to the factorization is the magnitude-compressed spectrogram, proposed by [10] and later utilized e.g. in [11, 5], which is defined as

$$\tilde{\mathbf{p}}_{ft} = \left[|P_{1ft}|^{1/2}\sigma(P_{1ft}), \ldots, |P_{Lft}|^{1/2}\sigma(P_{Lft})\right]^{\mathrm{T}}, \tag{3}$$

where $\sigma(C) = C/|C|$ is the signum function for any complex number C. This approach is adopted here since it ensures that the values on the main diagonal of the SCM consist of the mixture magnitude spectrum, while it additionally prevents certain observations from being unnecessarily enhanced. With this, the instantaneous covariance matrix, empirically determined as $\mathbf{R}_{ft} = \tilde{\mathbf{p}}_{ft}\tilde{\mathbf{p}}_{ft}^{\mathrm{H}}$, can then be modelled as

$$\mathbf{R}_{ft} \approx \hat{\mathbf{R}}_{ft} = \sum_{j}^{J} \tilde{S}_{jfn}\hat{\mathbf{R}}_j = \sum_{j}^{J} \tilde{S}_{jfn}\mathbf{y}_j\mathbf{y}_j^{\mathrm{H}}, \tag{4}$$

with source magnitude spectrum defined as $\tilde{S}_{jfn} = (S_{jfn}S_{jfn}^{*})^{1/2}$.

Considering that more than a single anechoic steering vector is required to accurately represent a reverberant soundfield, the model of the SCM is defined as a weighted combination of multiple anechoic steering vectors $\mathbf{y}_d$, for a sufficiently high number of directions D distributed uniformly on the surface of a unit sphere. This can be written as

$$\mathbf{R}_j \approx \hat{\mathbf{R}}_j = \sum_{d}^{D} Z_{jd}\, \mathbf{R}_d = \sum_{d}^{D} Z_{jd}\, \mathbf{y}_d\, \mathbf{y}_d^{\mathrm{H}}, \tag{5}$$

where the weights Z_{jd}, also referred to as the spatial selector, occupy the range between 0 and 1 and are defined for each source separately.

The magnitude spectrograms are factorized using the well-known NTF model, which is given by [12]

$$\tilde{S}_{jfn} \approx \sum_{k}^{K} Q_{jk}W_{fk}H_{kt}, \tag{6}$$

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

Substitution of the source SCM model given by (5) and the spectrogram model given by (6) into (4) results in the final model for the NTF with SCM for Ambisonic signals

$$\hat{\mathbf{R}}_{ft} = \sum_{j}^{J}\sum_{k}^{K} Q_{jk}W_{fk}H_{kt}\sum_{d}^{D} Z_{jd}\mathbf{R}_d\ . \tag{7}$$

The model parameters $\Theta = \{Q_{jk}, W_{fk}, H_{kt}, Z_{jd}\}$ are estimated by minimizing the following negative log-likelihood

$$-\log\left(p\left(\mathbf{R}_{ft}|\Theta\right)\right) = \sum_{f}^{F}\sum_{t}^{T}\|\hat{\mathbf{R}}_{ft} - \mathbf{R}_{ft}\|_{\mathrm{F}}^2, \tag{8}$$

which corresponds to maximization of the likelihood for the following probabilistic model

$$p\left(\mathbf{R}_{ft}|\Theta\right) = \prod_{f,t,a,b}^{F,T,A,B}\mathcal{N}_c\left([\hat{\mathbf{R}}_{ft}]_{ab}|[\mathbf{R}_{ft}]_{ab}, 1\right), \tag{9}$$

where $\mathcal{N}_c$ denotes the complex Gaussian distribution.

2.3. Incorporation of the Wishart Localization Prior

As the spatial selector Z_{jd} initialization strategy proposed in [5] is based on the estimation or an a priori knowledge of the DoAs corresponding to the source locations, we propose the following modifications to optimally exploit this information.

We follow the formulations presented in [13], where different localization priors for reverberant sound source separation were proposed, and we select the Wishart localization prior due to its inherent property to constrain the SCM and its effective performance in the sound source separation task. The Wishart probability density distribution over a Hermitian positive definite matrix $\mathbf{R}$ is defined as

$$\mathcal{W}\left(\mathbf{R}|\mathbf{\Psi}, \nu\right) = \frac{|\mathbf{\Psi}|^{-\nu}|\mathbf{R}|^{(\nu-L)}e^{-\mathrm{tr}\left(\mathbf{\Psi}^{-1}\mathbf{R}\right)}}{\pi^{L(L-1)/2}\prod_{l}^{L}\Gamma\left(\nu-l+1\right)}, \tag{10}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator, $\mathbf{\Psi}$ is the scale matrix, ν stands for the degrees of freedom and $\Gamma(\cdot)$ is the gamma function. The hyper-parameter ν allows to control the acceptable deviation from the mean, while the density, its mean and variance are finite for $\nu > L-1$, $\nu > L$ and $\nu > L+1$, respectively. Note that ν does not necessarily have to be an integer.

We next formulate the localization prior through the constraint on the individual source SCMs, by substituting $\mathbf{R}$ with $\hat{\mathbf{R}}_j$ in the probability density distribution (10). Considering that the distribution mean is equal to $\nu\mathbf{\Psi}$, we set the scale matrix to be $\mathbf{\Psi}_j = \frac{1}{\nu}\left(\mathbf{y}_j\mathbf{y}_j^{\mathrm{H}} + \epsilon\mathbf{I}\right)$, based on the steering vectors for the known DoAs $\mathbf{y}_j$ and the relative strength of the diffuse component ϵ, for which the covariance matrix is simply an identity matrix $\mathbf{I}$. Consequently, we derive the posterior probability as

$$p\left(\Theta|\mathbf{R}_{ft}\right) = \prod_{j,f,t,a,b}^{J,F,T,A,B} \mathcal{N}_c\left([\hat{\mathbf{R}}_{ft}]_{ab}|[\mathbf{R}_{ft}]_{ab}, 1\right) \mathcal{W}\left(\hat{\mathbf{R}}_j|\mathbf{\Psi}_j, \nu\right) \tag{11}$$

and the corresponding negative log-posterior

$$-\log\left(p\left(\Theta|\mathbf{R}_{ft}\right)\right) \overset{\text{c}}{=} \sum_{f,t}^{FT} \|\hat{\mathbf{R}}_{ft} - \mathbf{R}_{ft}\|_{\text{F}}^2 + \\ + \sum_{j}^{J} \left[\nu \text{tr}\left(\mathbf{\Psi}_j^{-1}\hat{\mathbf{R}}_j\right) - (\nu - L)\log\left|\hat{\mathbf{R}}_j\right|\right], \tag{12}$$

with the constant terms omitted for brevity in (12). Minimization of the proposed extended negative log-posterior (12) allows to estimate the model parameters Θ, taking into consideration the prior knowledge of the DoAs associated with source location.

2.4. Derivation of update equations for the proposed algorithm

Similarly to [10], the negative log-posterior (12) can be indirectly optimized using the so-called majorization scheme [14, 15], where a complicated minimization problem is simplified by defining suitable latent components and a corresponding auxiliary function. In this work, we choose the following latent components

$$\mathbf{C}_{ftjkd} = \hat{\mathbf{R}}_{ft}^{-1} Q_{jk} W_{fk} H_{kt} Z_{jd} \mathbf{R}_d, \tag{13}$$

such that the following constraint on the Hermitian positive definite matrix $\mathbf{C}_{ftjkd}$ is satisfied $\sum_{j,k,d}^{J,K,D} \mathbf{C}_{ftjkd} = \mathbf{I}$. Next, we define the appropriate auxiliary function

$$\mathcal{L}^+\left(\Theta, \mathbf{C}\right) = \sum_{f,t,j,k,d}^{F,T,J,K,D} Q_{jk}^2 W_{fk}^2 H_{kt}^2 Z_{jd}^2 \text{ tr}\left(\mathbf{R}_d \mathbf{C}_{ftjkd}^{-1} \mathbf{R}_d\right) - \\ 2\sum_{f,t,j,k,d}^{F,T,J,K,D} Q_{jk} W_{fk} H_{kt} Z_{jd} \text{ tr}\left(\mathbf{R}_{ft}\mathbf{R}_d\right) + \\ \sum_{j}^{J} \left[\nu \text{ tr}\left(\mathbf{\Psi}_j^{-1}\hat{\mathbf{R}}_j\right) - (\nu - L)\log\left|\hat{\mathbf{R}}_j\right|\right]. \tag{14}$$

By following the approach presented in [10] it can be shown that minimization of the auxiliary function (14) leads to an indirect minimization of the negative log-posterior (12). The optimization with respect to the model parameters Θ is performed by calculating partial derivatives of (14), such that in terms of Z_{jd} it is given by

$$\frac{\partial \mathcal{L}^+\left(\Theta, \mathbf{C}\right)}{\partial Z_{jd}} = 2\sum_{f,t,k}^{F,T,K} Q_{jk}^2 W_{fk}^2 H_{kt}^2 Z_{jd} \text{ tr}\left(\mathbf{R}_d \mathbf{C}_{ftjkd}^{-1} \mathbf{R}_d\right) - \\ 2\sum_{f,t,k}^{F,T,K} Q_{jk} W_{fk} H_{kt} \text{ tr}\left(\mathbf{R}_{ft}\mathbf{R}_d\right) + \\ \nu \text{ tr}\left(\mathbf{\Psi}_j^{-1}\mathbf{R}_d\right) - (\nu - L)\text{ tr}\left(\hat{\mathbf{R}}_j^{-1}\mathbf{R}_d\right). \tag{15}$$

The Multiplicative Update rules are obtained by substituting $\mathbf{R}_d\mathbf{C}_{ftjkd}^{-1}\mathbf{R}_d = Q_{jk}^{-1}W_{fk}^{-1}H_{kt}^{-1}Z_{jd}^{-1}\hat{\mathbf{R}}_{ft}\mathbf{R}_d$ in (15) and applying the gradient descent technique with appropriate adaptive step [16]. This enables to derive the iterative update equations according to the MU rule

$$\alpha \leftarrow \alpha\frac{[\nabla_\alpha f(\alpha)]_-}{[\nabla_\alpha f(\alpha)]_+}, \tag{16}$$

in which the updated value of a parameter α is calculated by multiplying its current value with the ratio of the negative to positive part of the gradient of a function $f(\alpha)$. The application of (16) to (15) results in the following update equation for the spatial selector

$$Z_{jd} \leftarrow Z_{jd}\left[\sum_{f,t,k}^{F,T,K} Q_{jk}W_{fk}H_{kt}\text{tr}\left(\mathbf{R}_{ft}\mathbf{R}_d\right) + \frac{\nu - L}{2}\text{tr}\left(\hat{\mathbf{R}}_j^{-1}\mathbf{R}_d\right)\right]\left[\sum_{f,t,k}^{F,T,K} Q_{jk}W_{fk}H_{kt}\text{tr}\left(\hat{\mathbf{R}}_{ft}\mathbf{R}_d\right) + \frac{\nu}{2}\text{tr}\left(\mathbf{\Psi}_j^{-1}\mathbf{R}_d\right)\right]^{-1}. \tag{17}$$

After each above Z_{jd} update the spatial selector weights are normalized to ensure that $\sum_d^D Z_{jd} = 1$.

On the other hand, the application of (16) to (14) with respect to Q_{jk}, W_{fk} and H_{kt} yields the MU equations that are equivalent to those provided in [5], which read

$$Q_{jk} \leftarrow Q_{jk}\frac{\sum_{f,t}^{F,T} W_{fk}H_{kt}\text{tr}\left(\mathbf{R}_{ft}\hat{\mathbf{R}}_j\right)}{\sum_{f,t}^{F,T} W_{fk}H_{kt}\text{tr}\left(\hat{\mathbf{R}}_{ft}\hat{\mathbf{R}}_j\right)}, \tag{18}$$

$$W_{fk} \leftarrow W_{fk}\frac{\sum_{j,k}^{J,K} Q_{jk}H_{kt}\text{tr}\left(\mathbf{R}_{ft}\hat{\mathbf{R}}_j\right)}{\sum_{j,k}^{J,K} Q_{jk}H_{kt}\text{tr}\left(\hat{\mathbf{R}}_{ft}\hat{\mathbf{R}}_j\right)}, \tag{19}$$

$$H_{kt} \leftarrow H_{kt}\frac{\sum_{f,j}^{F,J} Q_{jk}W_{fk}\text{tr}\left(\mathbf{R}_{ft}\hat{\mathbf{R}}_j\right)}{\sum_{f,j}^{F,J} Q_{jk}W_{fk}\text{tr}\left(\hat{\mathbf{R}}_{ft}\hat{\mathbf{R}}_j\right)}. \tag{20}$$

3. EXPERIMENTAL EVALUATION

3.1. Experiments and evaluation measures

The experimental evaluation was performed using the first-order Ambisonic signals as it is the most popular audio format based on the SHD representation. Nevertheless, the aforementioned derivations are universal for any SH order and in general as the order increases, the performance improves due to an enhanced spatial selectivity.

The test dataset was generated by convolving the anechoic samples of source signals with room impulse responses simulated using the image-source method [17]. A microphone array and two simultaneously active sound sources, were located inside a room with reverberation time of around 250 ms. Their positions were selected at random, while preserving the minimum angular separation between the two sources equal to 45° and keeping the distance to the microphone array of approximately 1.5 - 2 m. Two types of source signals were considered, namely the audio samples of musical instruments and speech utterances. The former include various types of instruments such as saxophone, violin, cello, guitar, bass and alto, which were taken from [18], while in the case of speech recordings, the utterances of distinct speakers from [19] were used.

We compared the performance of the proposed SH-SCM-NTF-WLP algorithm against the original SH-SCM-NTF algorithm [5] and state-of-the-art SH-GEM-GLP [1], which is a recently introduced SHD EM-based sound source separation method, that also incorporates a priori DoA information through the Gaussian Localization Prior. The evaluation was based on measures widely used for source separation, namely the signal-to-distortion ratio (SDR), the image-to-spatial-distortion-ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artefacts ratio (SAR) [20]. Each presented result is an average over 40 samples, for randomly selected signals of the different sources located at random positions.

All of the matrices in the considered NTF-based algorithms were initialized with the same random values, assuming that perfect DoA knowledge is available. In the proposed SH-SCM-NTF-WLP, the initial values of spatial selector Z_{jd} were uniformly set to 1, while the reference methods SH-SCM-NTF and SH-GEM-GLP were initialized according to the guidelines presented in the original papers [5] and [1], respectively.

In order to test the robustness of the proposed solution against deteriorating a priori localization, an additional experiment was performed in which the entire test dataset was repeated for progressively more corrupted DoAs.

In case of all performed experiments, the multichannel images of the source signals were reconstructed using the multichannel Wiener filter formulated in the Spherical Harmonic Domain, similar to that employed in [2].

3.2. Results and discussion

The results of the experimental evaluation for musical instruments and speech signals are presented in Fig. 1. They indicate that incorporation of the Wishart Localization Prior into the existing SH-SCM-NTF algorithm indeed brings substantial improvement in terms of the majority of evaluation measures. In particular, the proposed SH-SCM-NTF-WLP generally provides the highest overall separation quality with respect to the SH-SCM-NTF and the SH-GEM-GLP, as it reaches by far the highest scores in the SDR metric. In case of both the instrumental and the speech source signals, significant improvements are achieved within the ISR and the SIR measures, while maintaining roughly the same SAR. This suggest that by allowing more directions to contribute to the SCM model, the sound field is modelled more accurately, which enables to achieve more attenuation of the undesired sound source and better extraction of the desired signal, as given by the substantial improvements in the SIR measure. On the other hand, the improved separation and better preservation of the spatial properties, as given by the ISR, are not compromised by an introduction of notable artifacts, as confirmed by the SAR score which remains at a constant level for musical instruments and is only slightly reduced in case of the speech signals.

Figure 2 depicts the comparison of the SDR measures in case of deteriorating a priori localization, i.e. as a function of the increasing average angular error, for the original SH-SCM-NTF and the proposed SH-SCM-NTF-WLP. These results clearly indicate that although both methods exhibit somewhat similar rate of performance degradation, the proposed solution has an immense starting advantage, and thus can deliver more than satisfactory separation, even in case of high a priori localization error. Interestingly, the proposed SH-SCM-NTF-WLP method, even at the highest considered angular errors, still outperforms the original SH-SCM-NTF without any localization errors, in both instrumental and speech scenarios.

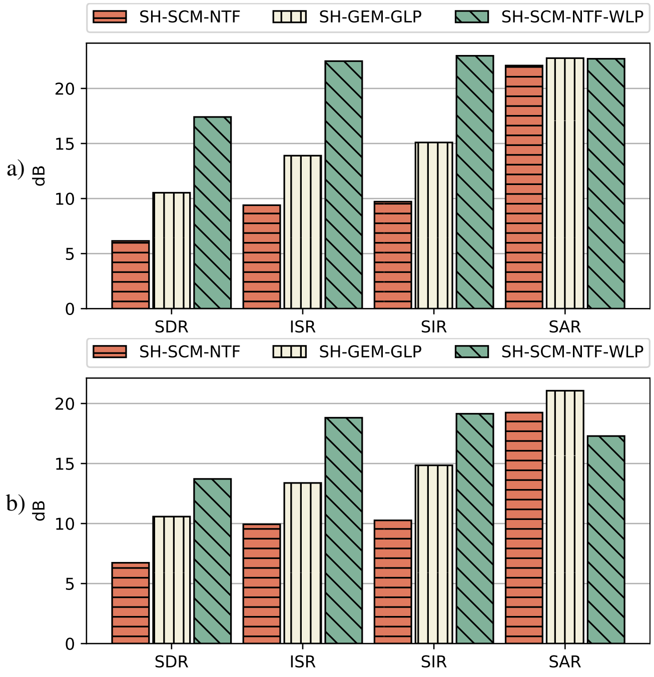

Fig. 1. The SDR, ISR, SIR, and SAR evaluation measures for the scenarios with (a) musical instruments and (b) the speech signals.

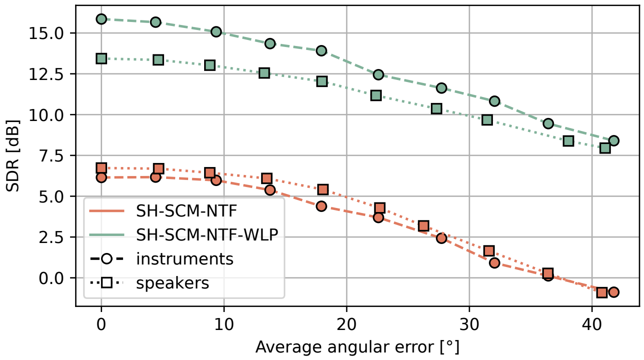

Fig. 2. The SDR measure as a function of the increasing average angular localization error, for the proposed SH-SCM-NTF-WLP and the original SH-SCM-NTF methods in both instrumental and speech scenarios.

4. CONCLUSIONS

In this work, we have proposed an extension of the existing Non-negative Tensor Factorization based algorithm for Ambisonic sound source separation in reverberant conditions. In particular, we introduced a constraint on the Spherical Harmonic Domain Spatial Covariance Matrix model, using the Wishart distribution and consequently we formulated a suitable posterior probability. As a result, we derived the extended Multiplicative Update rules for the estimation of models parameters. The evaluation performed for the first-order Ambisonic recordings of musical instruments and speech utterances shows improvement in source separation over the original and other reference methods.

5. REFERENCES

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

[2] Yuki Mitsufuji, Norihiro Takamune, Shoichi Koyama, and Hiroshi 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.

[3] Mohammed Hafsati, Nicolas Epain, Rémi Gribonval, and Nancy Bertin, "Sound source separation in the higher order ambisonics domain," in *DAFx 2019-22nd International Conference on Digital Audio Effects*, 2019, pp. 1–7.

[4] Yann Salaün, Emmanuel Vincent, Nancy Bertin, Nathan Souviraa-Labastie, Xabier Jaureguiberry, Dung T Tran, and Frédéric Bimbot, "The flexible audio source separation toolbox version 2.0," in *ICASSP*, 2014.

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

[6] Sachin N Kalkur, Sandeep Reddy C, and Rajesh 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.

[7] Vishnuvardhan Varanasi and Rajesh 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.

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

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

[10] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori 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.

[11] Joonas Nikunen and Tuomas Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 3, pp. 727–739, 2014.

[12] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shunichi Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons, 2009.

[13] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 1–11, 2013.

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

[15] Albert W Marshall, Ingram Olkin, and Barry C Arnold, *Inequalities: theory of majorization and its applications*, vol. 143, Springer, 1979.

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

[17] Jont B Allen and David 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.

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

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

[20] Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian 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.