Time-Deconvolutive CNMF for Multichannel Blind Source Separation

Thadeu Luiz Barbosa Dias, Wallace Alves Martins, and Luiz Wagner Pereira Biscainho

***Abstract*— This paper tackles multichannel separation of convolutive mixtures of audio sources by using complex-valued non-negative matrix factorization (CNMF). We extend models proposed by previous works and show that one may tailor advanced single-channel NMF techniques, such as the deconvolutive NMF, to the multichannel factorization scheme. Additionally, we propose a regularized cost function that enables the user to control the distribution of the estimated parameters without significantly increasing the underlying computational cost. We also develop an optimization framework compatible with previous related works. Our simulations show that the proposed deconvolutive model offers advantages when compared to the simple NMF, and that the regularization is able to steer the parameters towards a solution with desirable properties.**

***Keywords*— Blind source separation, convolutive mixture, NMF, deconvolutive NMF**

I. Introduction

Source separation has several applications: a celebrated example is the use of independent component analysis (ICA) to separate muscular activity interference from brain activity in encephalographic scans [1]; another interesting example is the use of BSS in speech enhancement for hearing aid devices [2].

Among traditional techniques for source separation, non-negative matrix factorization (NMF), a single-channel method, has been successfully employed in literature [3]. NMF factorizes the input matrix comprising non-negative entries as two smaller matrices, and is able to extract the most significant components that *explain* the observed data, i.e., a model and a set of parameters that produce a satisfactory estimate of the data. A distinction when compared to other rank-reducing methods, such as singular-value decomposition (SVD), is that the extracted components are themselves composed of non-negative entries: this is key for feature extraction when the features are non-negative by nature, which is the case for magnitude or power spectrograms.

In the source separation scenario, the resulting NMF factors from a mixture spectrogram can be thought of as a set of spectral signatures and temporal activation patterns [4]. It is expected that subsets of the extracted signatures explain each source, and a typical challenge is how to assign which components belong to each source. Usually, this assignment relies on some other prior information.

Mr. Dias is with the Electrical Engineering Program (PEE/Coppe) of Federal University of Rio de Janeiro (UFRJ); Prof. Martins is with the Department of Electronics and Computer Engineering (DEL/Poli) & PEE/Coppe, UFRJ (on leave), and with the University of Luxembourg (postdoc); Prof. Biscainho is with DEL/Poli & PEE/Coppe, UFRJ. E-mails: {thadeu.dias, wallace.martins, wagner}@smt.ufrj.br. This work was partially supported by Capes (88882.331631/2019-01), CNPq (PQ 306331/2017-9), and Faperj.

A more powerful way to realize source separation is to exploit spatial properties, as in the case of multichannel processing methods. A development by [5], named complex-valued NMF (CNMF), is the introduction of Hermitian positive semidefinite matrices constructed from the complex spectrograms as data points. Building on this model, the authors in [6] propose a geometric constraint on the CNMF parameters, providing spatially-coherent factorization to enhance the separation quality of the method.

This paper shows that the deconvolutive NMF model [7] can be tailored to the CNMF framework with good results, and that we may regularize the related cost function towards a sparse solution. We describe the signal representation in Section II, and the constrained construction of channel matrices as well as the application of the deconvolutive model in Section III. We derive the estimation framework for an Euclidean cost function in Section IV, and present the results in Section V. Finally, we briefly discuss our results and future works in Section VI.

II. Signal Representation

Considering an array with M sensors, and the propagation media as linear time-invariant, the convolutive mixture as acquired by the individual sensors can be written as

$$x_m(t) = \sum_{q=1}^{Q} \int_{-\infty}^{+\infty} h_{qm}(t-\tau) s_q(\tau)\,\mathrm{d}\tau, \tag{1}$$

where Q is the true number of sources, $x_m(t)$ is the m^{th} sensor measurement, $h_{qm}(t)$ is the impulse response relative to the channel between the source-sensor pair (q,m), and $s_q(t)$ is the true emission of source q. Translating this relationship to the short-time Fourier transform (STFT) domain, each (complex) time-frequency point measurement x_{ilm} is

$$x_{ilm} = \sum_{q=1}^{Q} h_{iqm} s_{ilq}, \tag{2}$$

where i denotes the frequency bin, l is an index for the time frame, h_{iqm} is the frequency response at bin i of the channel relative to the source-sensor pair (q,m), and s_{ilq} is the STFT of the emission of source q at time-frequency point (i,l).

In order to represent the overall measurements as Hermitian matrices, we take the outer product of the vector $\boldsymbol{x}_{il}$ formed by the measurements across all sensors in a single time-frequency point, forming the matrices

$$\boldsymbol{X}_{il} = \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \boldsymbol{h}_{iq} \boldsymbol{h}_{iq'}^{\mathrm{H}} s_{ilq} s_{ilq'}^{*}. \tag{3}$$

Considering uncorrelated sources, we may invoke uncorrelatedness between the STFT coefficients, $\forall q \neq q'$, $\mathbb{E}\left[s_{ilq} s^*_{ilq'}\right] = 0$, and neglect the crossed terms. Furthermore, with the intention of factorizing the magnitude spectra, as proposed in [5], the STFT coefficients are mapped through a magnitude square-root function: $\phi(z) = \frac{z}{\sqrt{|z|}}$. Then we can rewrite (3) as

$$\boldsymbol{X}_{il} \approx \sum_{q=1}^{Q} \boldsymbol{h}_{iq} \boldsymbol{h}_{iq}^{\mathrm{H}} |s_{ilq}|, \tag{4}$$

and define $\boldsymbol{H}_{iq} = \boldsymbol{h}_{iq}\boldsymbol{h}_{iq}^{\mathrm{H}}$ as a matrix that encodes the phase properties of source q at bin i. The entries of $\boldsymbol{H}_{iq}$ encode the phase difference between the responses of a channel pair. By the outer product construction, $\boldsymbol{H}_{iq}$ preserves phase information without actually modeling the absolute phase of the measurements. Expression (4) then motivates a joint factorization of the sources' magnitude spectra and spatial-property matrices.

III. Factorization model

The main idea is to explore the compressibility of the magnitude representation in order to find K components that best explain the measurements. In the context of CNMF, we seek to explain the measured data points $\boldsymbol{X}_{il}$ as non-negative combinations of positive semidefinite matrices. We assign to each NMF component a family of spatial-property matrices, and cluster components based on their spatial properties when reconstructing the sources. The overall model for the measured data points can be written as

$$\boldsymbol{X}_{il} \approx \sum_{k=1}^{K} \boldsymbol{H}_{ik} \check{s}_{ilk}, \tag{5}$$

where $\boldsymbol{H}_{ik}$ encodes spatial properties for a component and $\check{s}_{ilk}$ is a magnitude estimate computed through the NMF framework. In the following, we detail how the parameters are obtained.

A. Magnitude activation model

In the single channel case, the standard NMF finds a low rank approximation to some data matrix $\boldsymbol{S} \in \mathbb{R}_+^{I\times L}$ as the product of two smaller non-negative matrices $\boldsymbol{B} \in \mathbb{R}_+^{I\times K}$ and $\boldsymbol{G} \in \mathbb{R}_+^{K\times L}$, where, usually, $K \ll \mathrm{Rank}(\boldsymbol{S})$. If the input data matrix consists of a magnitude spectrogram, with bins as rows and time frames as columns, a useful interpretation of the extracted matrices arises: the columns $\boldsymbol{b}_k \in \mathbb{R}_+^I$ of $\boldsymbol{B}$ are spectral signatures present in the measurements, and the rows $\boldsymbol{g}_k \in \mathbb{R}_+^L$ of $\boldsymbol{G}$ are activation patterns for such signatures across time frames. The application of the simple NMF to the CNMF model would lead to the magnitude estimate $\check{s}_{ilk} = b_{ik} g_{kl}$. We propose the estimation of $\check{s}_{ilk}$ through a deconvolutive NMF model: the extracted signatures have a set length $T \geq 1$, being represented as small matrices $\boldsymbol{B}_k \in \mathbb{R}_+^{I\times T}$ such that the collection of spectral signatures is the tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}_+^{I\times T\times K}$. The magnitude estimates are obtained through multiplication of the sub-components $\boldsymbol{b}_{kt} \in \mathbb{R}_+^I$ of $\boldsymbol{\mathcal{B}}$ by time-shifted versions of $\boldsymbol{g}_k$, such that the instantaneous magnitude estimate due to the k^{th} component is

$$\check{s}_{ilk} = \sum_{t=1}^{T} b_{itk} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l, \tag{6}$$

where the shift operator $\overrightarrow{[\cdot]}^t$ is equivalent to a post multiplication by a subdiagonal shift matrix (after a shift of length t, the first t columns are filled with zeroes).

In effect, the standard form of the CNMF with the deconvolutive model can be written as

$$\hat{\boldsymbol{X}}_{il} = \sum_{k=1}^{K} \boldsymbol{H}_{ik} \sum_{t=1}^{T} b_{itk} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l, \tag{7}$$

which shares the single channel deconvolutive NMF properties of being able to efficiently extract spectral patterns that vary with time. Considering that continuous emissions are a property present in many audio signals, this model is appropriate when moving to a more powerful separation model.

B. Spatial covariance model

We apply the direction-of-arrival (DoA) based factorization method introduced in [6] to the channel matrices $\boldsymbol{H}_{ik}$. An issue with the unconstrained estimation of matrices $\boldsymbol{H}_{ik}$ is that there is no guarantee that the set of matrices $\boldsymbol{\mathcal{H}}_k$ (all matrices $\boldsymbol{H}_{ik}$ with fixed k) actually encodes a single coherent single-input multiple-output channel between some component and the sensor array. Instead, the set $\boldsymbol{\mathcal{H}}_k$ is constructed as a non-negative linear combination of geometrically-defined beamforming kernel matrices $\boldsymbol{W}_{io}$.

Consider the scheme depicted in Fig. 1: with a sufficiently far emission source somewhere along the direction of $\boldsymbol{k}_o$ (a unit length vector), such that the wavefronts can be considered planar, the difference in propagation length can be calculated through the inner product $\langle \boldsymbol{p}_{m'} - \boldsymbol{p}_m, \boldsymbol{k}_o \rangle$ so that the relative time-difference of arrival (TDoA) is simply $\tau_{mm'}(\boldsymbol{k}_o) = \frac{\langle \boldsymbol{p}_{m'} - \boldsymbol{p}_m, \boldsymbol{k}_o \rangle}{c}$, where c denotes the wave propagation velocity.

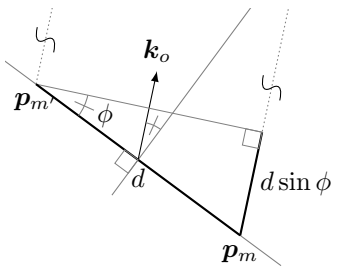

Fig. 1. TDoA as function of array geometry and wave incidence direction.

It is straightforward to find the frequency-dependent phase lag using Fourier transform properties, and from the non-normalized STFT bin frequencies, the per-bin phase lag can be calculated as

$$\theta_{mm'}(f_i, \boldsymbol{k}_o) = -2\pi f_i \tau_{mm'}(\boldsymbol{k}_o). \tag{8}$$

The idea for modelling $\boldsymbol{\mathcal{H}}$ is to sample O directions from the unit sphere around the array and form the beamforming kernels $\boldsymbol{W}_{io}$ for all STFT bins for each DoA sample. The beamforming kernels are $M \times M$ Hermitian matrices containing the relative phase shifts (for a set frequency f_i and direction $\boldsymbol{k}_o$) as complex factors:

$$\boldsymbol{W}_{io} = \begin{bmatrix} 1 & \mathrm{e}^{\mathrm{j}\theta_{12}(f_i,\boldsymbol{k}_o)} & \dots & \mathrm{e}^{\mathrm{j}\theta_{1M}(f_i,\boldsymbol{k}_o)} \\ \mathrm{e}^{\mathrm{j}\theta_{21}(f_i,\boldsymbol{k}_o)} & 1 & \dots & \mathrm{e}^{\mathrm{j}\theta_{2M}(f_i,\boldsymbol{k}_o)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{e}^{\mathrm{j}\theta_{M1}(f_i,\boldsymbol{k}_o)} & \mathrm{e}^{\mathrm{j}\theta_{M2}(f_i,\boldsymbol{k}_o)} & \dots & 1 \end{bmatrix}.$$

Finally, the channel matrices $\boldsymbol{H}_{ik}$ can be modeled in terms of the DoA kernels as non-negative linear combinations

$$\boldsymbol{H}_{ik} = \sum_{o=1}^{O} z_{ko} \boldsymbol{W}_{io}, \tag{9}$$

where the factors $z_{ko} \in \mathbb{R}_+$ are shared across all frequencies for a given component, making this a spatially coherent factorization. Additionally, this model allows the spatial properties encoded by the vectors $\boldsymbol{z}_k$ to be clustered, since it is expected that components with similar spatial signatures belong to the same source.

C. Complete model

The complete model for the measured covariance matrices in terms of deconvolutive CNMF parameters can be written as

$$\hat{\boldsymbol{X}}_{il} = \sum_{k=1}^{K} \sum_{o=1}^{O} z_{ko} \boldsymbol{W}_{io} \sum_{t=1}^{T} b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l \,. \tag{10}$$

This factorization is unique up to scaling factors, so additional constraints are used, namely $\sum_o z_{ko}^2 = 1$, $\sum_l g_{kl}^2 = 1$, and $\|\boldsymbol{W}_{io}\|_{\mathrm{F}} = 1$, where $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A})}$ denotes the Frobenius norm of a matrix. In order to find the best estimates for the parameters, a statistical model can be assigned, and an estimation framework can be structured.

D. Source reconstruction

Given the CNMF parameter estimates, the per-source spectral images can be reconstructed through a Wiener filter

$$\boldsymbol{y}_{ilq} = \boldsymbol{x}_{il} \frac{\sum_{k,o} \beta_{qk} z_{ko} \sum_t b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l}{\sum_{q,k,o} \beta_{qk} z_{ko} \sum_t b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l}, \tag{11}$$

where β_{qk} are learned membership coefficients relating a given component k with source q. The membership coefficients can be obtained through a regular clustering algorithm, such as k-means, c-means, or even NMF (considering that the spatial factors z_{ko} are also non-negative). The overall effect is to multiply the input spectrograms with a mask of ratios of estimated spectral magnitudes, preserving the original phase.

IV. Parameter estimation

Following previous works [5], [6], the considered generative model for the entries of the data matrices $\boldsymbol{X}_{il}$ are that of independent complex Gaussian variables with unit variance. What follows is that the maximum likelihood estimate is obtained through a squared error minimization problem. In fact, the likelihood function for the parameters, considering the overall measurement can be written as

$$\mathcal{L}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \propto \prod_{i=1}^{I} \prod_{l=1}^{L} \exp\left(-\frac{\|\boldsymbol{X}_{il} - \hat{\boldsymbol{X}}_{il}\|_{\mathrm{F}}^2}{2} \right). \tag{12}$$

It can be useful to embed some prior on the parameters. This knowledge can be directly related to a regularization factor, steering the algorithm towards a solution with some desirable properties. In this paper we consider the generative model for the spectral signatures b_{ikt} as a one-sided exponential distribution with some scaling factor $\alpha_{\boldsymbol{\mathcal{B}}}$, leading to the regularized likelihood

$$\mathcal{L}_{\mathrm{R}} \propto \exp\left(-\alpha_{\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_1\right) \mathcal{L}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \tag{13}$$

and regularized cost function

$$\ell_{\mathrm{R}}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) = 2\alpha_{\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_1 + \sum_{i=1}^{I} \sum_{l=1}^{L} \|\boldsymbol{X}_{il} - \hat{\boldsymbol{X}}_{il}\|_{\mathrm{F}}^2, \tag{14}$$

corresponding to the negative log-likelihood function (with omitted constants and rescaled for convenience), where the tensor ℓ_1-norm is defined as $\sum_{i,k,t} b_{ikt}$. This is inspired by a LASSO [8] regression, where selectiveness of the signatures $\boldsymbol{B}_k$ (consequently, a more sparse representation) is desired.

A. Minimization procedure

While (14) is non-convex relative to the CNMF parameters, it is individually convex over $\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}$, and $\boldsymbol{G}$. Thus, a block relaxation minimization scheme [9] to obtain good solutions may be employed. We consider an auxiliary function to (14), namely:

$$\ell_{\mathrm{R}}^{+} = 2\alpha_{\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_1 + \sum_{i,l,k,o,t} \frac{1}{r_{ilkot}} \|\boldsymbol{S}_{ilkot} - z_{ko} \boldsymbol{W}_{io} b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l\|_{\mathrm{F}}^2, \tag{15}$$

where r_{ilkot} are any positive variables satisfying $\sum_{k,o,t} r_{ilkot} = 1$, and $\boldsymbol{S}_{ilkot}$ are Hermitian matrices satisfying $\sum_{k,o,t} \boldsymbol{S}_{ilkot} = \boldsymbol{X}_{il}$. It can be proven that

$$\ell_{\mathrm{R}}^{+}(\boldsymbol{\mathcal{S}}, \boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \geq \ell_{\mathrm{R}}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}), \quad \text{and} \tag{16}$$

$$\min_{\boldsymbol{\mathcal{S}}} \ell_{\mathrm{R}}^{+}(\boldsymbol{\mathcal{S}}, \boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) = \ell_{\mathrm{R}}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}). \tag{17}$$

Through a constrained minimization using Lagrange multipliers, the optimal values for $\boldsymbol{S}_{ilkot}$ can be derived as

$$\boldsymbol{S}_{ilkot}^{*} = z_{ko} \boldsymbol{W}_{io} b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l - r_{ilkot} \boldsymbol{E}_{il}, \tag{18}$$

where $\boldsymbol{E}_{il}$ is the error matrix $\boldsymbol{X}_{il} - \hat{\boldsymbol{X}}_{il}$. Combined with the majorizer conditions (16) and (17), individual minimization of (15) across the CNMF variables with $\boldsymbol{\mathcal{S}}$ set as $\boldsymbol{\mathcal{S}}^{*}$ is guaranteed non-increasing. With the auxiliary definition

$$\hat{x}_{il} = \sum_{k,o,t} z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l, \tag{19}$$

a useful way to define r_{ilkot} arises, namely

$$r_{ilkot} = \frac{z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}{\hat{x}_{il}}. \tag{20}$$

Although this definition is not strictly positive, it is safe to ignore the zeroed values, since

$$\lim_{r_{ilkot} \to 0} \frac{1}{r_{ilkot}} \| \boldsymbol{S}^*_{ilkot} - z_{ko} \boldsymbol{W}_{io} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l \|^2_{\mathrm{F}} = 0,$$

and this definition allows for implicit computation of the $\boldsymbol{S}^*_{ilkot}$ factors.

Replacing $\boldsymbol{\mathcal{S}}^*$ and r_{ilkot} with their definitions, the multiplicative rules for non-negative factors can be obtained:

$$z_{ko} \longleftarrow z_{ko} \left[\frac{\sum_{i,l,t} \left(\hat{x}_{il} + \mathrm{tr}\left(\boldsymbol{E}_{il} \boldsymbol{W}_{io} \right) \right) b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}{\sum_{i,l,t} \hat{x}_{il} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l} \right], \tag{21}$$

$$b_{ikt} \longleftarrow b_{ikt} \left[\frac{\sum_{l,o} \left(\hat{x}_{il} + \mathrm{tr}\left(\boldsymbol{E}_{il} \boldsymbol{W}_{io} \right) \right) z_{ko} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}{\alpha_{\boldsymbol{\mathcal{B}}} + \sum_{l,o} \hat{x}_{il} z_{ko} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l} \right], \tag{22}$$

$$g_{kl} \longleftarrow g_{kl} \left[\frac{\sum_{i,o,t} ([\overleftarrow{\hat{\boldsymbol{x}}_i}^{t-1}]_l + \mathrm{tr}([\overleftarrow{\boldsymbol{\mathcal{E}}_i}^{t-1}]_l \boldsymbol{W}_{io})) z_{ko} b_{ikt}}{\sum_{i,o,t} [\overleftarrow{\hat{\boldsymbol{x}}_i}^{t-1}]_l z_{ko} b_{ikt}} \right]. \tag{23}$$

The update process for the kernel matrices is slightly different, as only optimization of the magnitudes are allowed, and the positive semidefinite constraint must be accounted for. What follows is an optimization scheme similar to a projected gradient algorithm: the possibly unfeasible point that minimizes the cost function is calculated as

$$\hat{\boldsymbol{W}}_{io} \longleftarrow \frac{\sum_{l,k,t} z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l [\hat{x}_{il} \boldsymbol{W}_{io} + \boldsymbol{E}_{il}]}{\sum_{l,k,t} \hat{x}_{il} z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}; \tag{24}$$

this point is projected onto the positive semidefinite space by rectification of its eigenvalues:

$$\boldsymbol{V}_{io} \boldsymbol{\Lambda}_{io} \boldsymbol{V}^{\mathrm{H}}_{io} \longleftarrow \hat{\boldsymbol{W}}_{io} \tag{25}$$

$$\hat{\boldsymbol{W}}^{+}_{io} \longleftarrow \boldsymbol{V}_{io} \boldsymbol{\Lambda}^{+}_{io} \boldsymbol{V}^{\mathrm{H}}_{io}; \tag{26}$$

finally only the entries' magnitudes are updated, as the true update is obtained as

$$\boldsymbol{W}_{io} \longleftarrow \mathrm{abs}(\hat{\boldsymbol{W}}^{+}_{io}) \odot \mathrm{sign}(\boldsymbol{W}_{io}), \tag{27}$$

in which both $\mathrm{abs}(\cdot)$ and $\mathrm{sign}(\cdot)$ operate elementwise on their arguments, and $\odot$ denotes the matrix Hadamard product. Concerning the scaling factors, after each update $\boldsymbol{z}_k$ and $\boldsymbol{g}_k$ are normalized to unity, while the reciprocal correction factor is applied to $\boldsymbol{B}_k$, that is:

$$v_k \longleftarrow \|\boldsymbol{z}_k\|_2 : \quad \boldsymbol{z}_k \longleftarrow \frac{\boldsymbol{z}_k}{v_k} \quad \boldsymbol{B}_k \longleftarrow v_k \boldsymbol{B}_k, \qquad \text{and}$$

$$v_k \longleftarrow \|\boldsymbol{g}_k\|_2 : \quad \boldsymbol{g}_k \longleftarrow \frac{\boldsymbol{g}_k}{v_k} \quad \boldsymbol{B}_k \longleftarrow v_k \boldsymbol{B}_k.$$

Similarly, $\boldsymbol{W}_{io}$ is rescaled to unity Frobenius norm, but no rescaling of other parameters is needed:

$$\boldsymbol{W}_{io} \longleftarrow \frac{\boldsymbol{W}_{io}}{\|\boldsymbol{W}_{io}\|_{\mathrm{F}}}.$$

V. Numerical results

The program developed to assess the performance of the proposals was coded in Python, using TensorfFlow v1.4, and executed on an Intel Xeon Gold 5120. We consider the separation of two sound sources positioned 90° apart, with two mixtures synchronously captured by omnidirectional microphones placed 8 cm apart from each other, as illustrated in Fig. 2. The considered tracks are two vocal samples, about 11 s long, sampled at 22.05 kHz, drawn from the DSD100 dataset [10]. A closed room with $\mathrm{RT}_{60} \approx 300$ ms was simulated using CATT-Acoustic v9.0c [11], and the two sources were positioned on the central horizontal plane of the room. The microphone pair was positioned at the center of the room. Since only two sensors were used, a geometric ambiguity arises, and it is enough to sample directions from any half-plane defined by the segment connecting the sensor pair. Thus, the DoA sampling scheme depicted in Fig. 2 was used, with $O = 60$ directions.

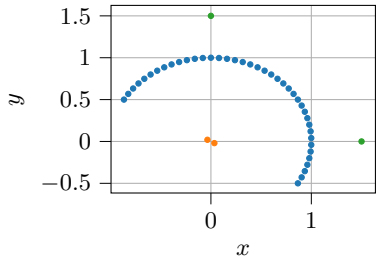

Fig. 2. DoA sampling and array geometry: sensor positions in orange, DoA in blue, and true sources in green.

Square-rooted Hanning windows were used for STFT analysis and synthesis, with 50% overlap. Frame length was chosen as 1024 samples, corresponding to approximately 46 ms, $I = 513$ bins, and $L = 490$ frames. We considered a factorization using $K = 60$ components, and deconvolutive length $T = 5$.

The algorithm ran for 500 iterations, with $\alpha_{\boldsymbol{\mathcal{B}}} = 0.5$. The DoA sampling scheme allows for an ordered indexing of the samples based on the angle w.r.t. the sensor axis, and the obtained spatial features $\boldsymbol{Z}$ are depicted in Fig. 3. Two fairly distinct clusters can be observed, with high activations around directions 50 and 20, corresponding to the directions East and North, respectively, in Fig. 2.

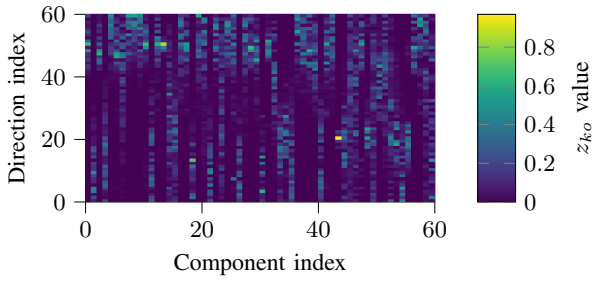

Fig. 3. Per-component spatial features. Most components have a localized DoA composition, corresponding to tight 'bouquets' around a particular DoA.

We performed weighted k-means on the vectors $\boldsymbol{z}_k$, with weights corresponding to the component energy $\|\boldsymbol{B}_k\|^2_{\mathrm{F}}$. The

source-component membership coefficients β_{qk} were set to 1 or 0 based on the obtained clustering, and the two centroids $\boldsymbol{z}_q$ corresponding to the averaged spatial properties for each source are depicted in Fig. 4.

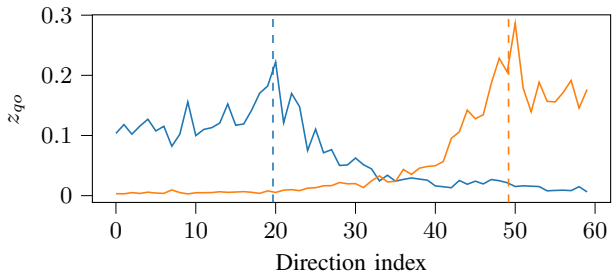

Fig. 4. $\boldsymbol{z}_q$ spatial features: cluster 1 in blue, and cluster 2 in orange. The true directions of the sources are depicted in dashed lines.

The centroids' peak activations closely match the true directions for each source, such that a rough estimate of the sources directions can be obtained from the method, although the estimate is likely to deteriorate in heavily reverberant environments.

The separation quality was measured using the `mir_eval` suite [12], and we benchmarked our results against a non-regularized, non-deconvolutive CNMF ($T = 1$ and $\alpha_{\boldsymbol{\mathcal{B}}} = 0$). The per-source signal-to-distortion ratio (SDR) scores for the proposed method were 7.36 dB and 4.08 dB, while the reference scored 6.17 dB and 2.92 dB, respectively. The evaluated source-to-interference ratios (SIR) for the proposed model were 15.03 dB and 4.66 dB; the reference scores were 11.93 dB and 3.20 dB, respectively.

The proposed method, therefore, outperformed the benchmark in the tested setups at the expense of a slight increase (measured around 30%) in computational time.

We additionally tested the effects of regularization on the energy distribution between components with three different $\alpha_{\boldsymbol{\mathcal{B}}}$ values. The histogram in Fig. 5 shows the sparsity-inducing effects, especially in the more strictly regularized case, effectively annihilating some components, which could be interpreted as an enforced selectiveness on the components.

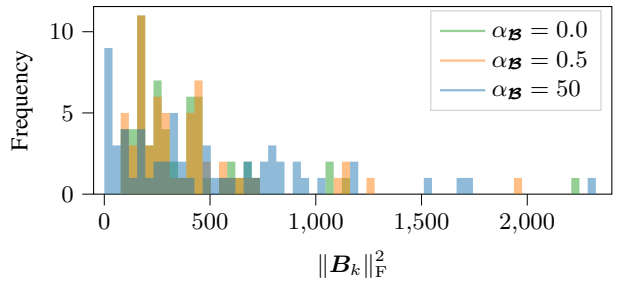

Fig. 5. Component energy distribution with different regularization values. Overregularization can be observed with $\alpha_{\boldsymbol{\mathcal{B}}} = 50$, as several components were reduced to insignificant energy.

In general terms, one would desire sparse signatures if the data has a sparse nature; this property is usually manifested in tonal emissions, where the emitted energy is well localized in the frequency domain. In this sense, it is expected that the imposed regularization enhances the algorithm's precision for tonal emissions, while some degradation can occur in percussive or other non tonal emissions.

VI. Conclusions

We proposed an extended version of the CNMF algorithm, leveraging the efficient representation from the deconvolutive NMF model. We also provided user with control on the distribution of the extracted signatures through regularization, which steers the method towards a sparse solution. Our proposed method is a generalization of the baseline algorithm, with added flexibility, able to efficiently factorize signals with complex spectral signature.

The simulations indicate that the proposed technique enjoys superior capabilities regarding the separation task, although a more extensive evaluation to explore the large number of hyperparameters is needed before drawing definitive conclusions.

Future works include the extension of these ideas to other NMF models, such as multi-layer or projective NMF, and different cost functions, such as Itakura-Saito divergence and other generalized $\alpha\beta$-divergences.

References

[1] M. Congedo, C. Gouy-Pailler, and C. Jutten, "On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics." *Clinical Neurophysiology*, vol. 119, no. 12, pp. 2677–2686, December 2008.

[2] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, March 2010, pp. 1–6.

[3] N. Gillis, "The why and how of nonnegative matrix factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines*, ser. Machine Learning and Pattern Recognition Series, J. A. K. Suykens, M. Signoretto, and A. Argyriou, Eds. Boca Raton, USA: Chapman & Hall/CRC, 2014, ch. 12, pp. 257–291.

[4] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, October 2003, pp. 177–180.

[5] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "New formulations and efficient algorithms for multichannel NMF," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Palz, USA, October 2011, pp. 153–156.

[6] J. Nikunen and T. 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, March 2014.

[7] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, C. G. Puntonet and A. Prieto, Eds. Berlin, Heidelberg: Springer, 2004, pp. 494–499.

[8] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[9] J. de Leeuw, "Block-relaxation algorithms in statistics," in *Information Systems and Data Analysis*, H.-H. Bock, W. Lenski, and M. M. Richter, Eds. Berlin, Heidelberg: Springer, 1994, pp. 308–324.

[10] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, Eds. Cham: Springer, August 2017, pp. 323–332.

[11] "CATT-Acoustic," (Visited on 10-Feb-2019). [Online]. Available: http://catt.se

[12] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "Mir_eval: A transparent implementation of common MIR metrics," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, October 2014, pp. 367–372.