Multichannel Source Separation Using Time-Deconvolutive CNMF

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

***Abstract*—This paper addresses the separation of audio sources from convolutive mixtures captured by a microphone array. We approach the problem using complex-valued non-negative matrix factorization (CNMF), and extend previous works by tailoring advanced (single-channel) NMF models, such as the deconvolutive NMF, to the multichannel factorization setup. Further, a sparsity-promoting scheme is proposed so that the underlying estimated parameters better fit the time-frequency properties inherent in some audio sources. The proposed parameter estimation framework is compatible with previous related works, and can be thought of as a step toward a more general method. We evaluate the resulting separation accuracy using a simulated acoustic scenario, and the tests confirm that the proposed algorithm provides superior separation quality when compared to a state-of-the-art benchmark. Finally, an analysis of the effects of the introduced regularization term shows that the solution is in fact steered toward a sparser representation.**

***Index Terms*—Blind source separation, convolutive mixture, NMF, deconvolutive NMF**

I. Introduction

Blind source separation (BSS) is an extensively researched topic with a wide variety of applications [2]. A celebrated example is the use of independent component analysis (ICA) [3, 4, 5] to separate muscular activity interference from brain activity in encephalographic scans [6]. Another interesting example is the use of BSS in speech enhancement for hearing aid devices [7].

Among traditional techniques for source separation, non-negative matrix factorization (NMF) [8], a single-channel method, has been successfully employed in the literature [9]. NMF factorizes an input matrix with non-negative entries into two lower-rank matrices with non-negative entries, 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. In comparison with other rank-reducing methods such as singular-value decomposition (SVD), the fact of dealing only with non-negative quantities is a distinctive feature of NMF: this is suitable for parameters that are non-negative by nature as is the case of magnitude or power spectra, and prevents mutual cancelling of components by destructive interference, since the model is strictly additive.

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 [10]. It is expected that subsets of the extracted signatures explain each source, and a typical challenge is how to assign which components correspond to each source. Usually, this assignment relies on some other prior information.

A more powerful way to perform source separation is to exploit spatial information, as in the case of multichannel processing methods. The complex-valued NMF (CNMF) [11] is a development in this direction, with the introduction of Hermitian positive semidefinite matrices as data points, derived from the measured signals' complex-valued spectrograms. Building on the CNMF model, an alternative factorization, with geometric constraints, is proposed in [12], providing spatially-coherent estimation, thus enhancing the separation quality of the method.

In [11, 12, 13], the factorization models were based on the standard NMF; by construction, the standard NMF model does not take into account the temporal sequence of samples, that is, a random shuffle in the time frames is irrelevant from the decomposition viewpoint. This approach, therefore, disregards useful continuity structures that can be observed in some data, such as musical samples, and often, the emission signatures cannot be efficiently represented by single NMF components. In order to address this limitation, we propose using a deconvolutive NMF (NMFD) scheme [14]. The extended NMFD signatures have a user-defined span of time frames, opening up the possibility for a more concise representation of some musical emissions, and eventually enhancing the separation quality.

In this paper we show that the deconvolutive NMF model can be tailored to the CNMF framework with good results, and that we may regularize the related cost function toward a sparse solution. We start with the problem statement in Section II, give a brief introduction to the NMF models in Section III, describe the transformations of the original data into the CNMF data points in Section IV, and the constrained construction of channel matrices as well as the application of the deconvolutive model in Section V. We derive the estimation framework for a Euclidean cost function in Section VI, and present the numerical results in Section VII.

A note on notation: In this paper, we denote scalars by regular lower-case letters, e.g. x, vector variables are lower-case bold letters, e.g. $\boldsymbol{x}$, matrices are upper-case bold, e.g. $\boldsymbol{X}$, and

This work is an extended version of our previous work [1], presented at the XXXVII Brazilian Symposium on Telecommunications and Signal Processing (SBrT-2019).

Mr. Thadeu L.B. Dias is with the Electrical Engineering Program (PEE/COPPE) of the Federal University of Rio de Janeiro (UFRJ); Prof. Wallace A. Martins is with the Department of Electronics and Computer Engineering (DEL/Poli) & PEE/COPPE, UFRJ (on leave), and with the Interdisciplinary Centre for Security, Reliability and Trust (SnT) of the University of Luxembourg; Prof. Luiz W.P. 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.

Digital Object Identifier: 10.14209/jcis.2020.11

higher-order tensors are calligraphic upper-case bold, e.g. $\boldsymbol{\mathcal{X}}$. An indexed element from a higher-order tensor corresponds to a slice of the original tensor, e.g. $[\boldsymbol{B}_k]_{it} = [\boldsymbol{\mathcal{B}}]_{ikt}$. Conversely, when some index is omitted and the variable gets promoted to a higher-order form, we refer to the entire collection of elements with the same name, that is, $\boldsymbol{\mathcal{W}} = \{\boldsymbol{W}_{io}\ \forall i, o\}$.

II. Problem Statement

We state our problem as that of separating meaningful acoustic sources captured by a microphone array with known geometry placed in a reverberant environment. Meaningful in this case means a source that is spatially concentrated, and emissions from different positions are considered as being from different sources. We consider that the propagation media is linear and time-invariant, and effects such as reverberation and multipath-propagation are then modeled by the unknown source-sensor channel impulse responses.

Under such considerations and assuming an array with M microphones, the acquired signals can then be described by the following model:

$$x_m(t) = \sum_{q=1}^{Q} (h_{qm} * s_q)(t), \tag{1}$$

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), $s_q(t)$ is the true emission of source q, and $*$ denotes the convolution operation. We would then like to be able to reconstruct estimates for the individual source images $\hat{y}_{qm}(t) = (h_{qm} * s_q)(t)$, as captured by the array.

We now describe the basic NMF formulations, and the steps toward integrating Eq. (1) into the non-negative framework.

III. NMF Basics

Non-negative matrix factorization is, at its core, the factorization of a matrix of non-negative entries into two reduced-rank matrices with non-negative entries.

In its most basic formulation [15], we have a data matrix $\boldsymbol{X} \in \mathbb{R}_+^{I \times L}$, and want to find a K-rank approximation of the original matrix according to a suitable criterion, subject to a non-negative constraint on the components; the approximation is given by matrices $\boldsymbol{B} \in \mathbb{R}_+^{I \times K}$ and $\boldsymbol{G} \in \mathbb{R}_+^{K \times L}$ such that

$$\boldsymbol{X} \approx \boldsymbol{B}\boldsymbol{G} = \sum_{k=1}^{K} \boldsymbol{b}_k \boldsymbol{g}_k^{\mathrm{T}}, \tag{2}$$

as illustrated in Fig. 1.

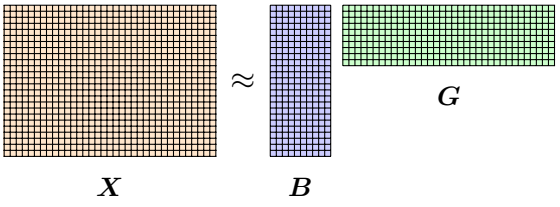

Fig. 1. Simple NMF model.

Usually, the approximation criterion is described by some cost function $\mathcal{L}(\boldsymbol{X}, \boldsymbol{B}\boldsymbol{G})$, and the factors are obtained by jointly minimizing the overall cost:

$$\begin{aligned} &\min_{\boldsymbol{B},\boldsymbol{G}} \ \mathcal{L}(\boldsymbol{X}, \boldsymbol{B}\boldsymbol{G}) \\ &\text{subject to non-negative constraints.} \end{aligned} \tag{3}$$

A useful interpretation of the extracted factors is the following: if the rows of $\boldsymbol{X}$ are features, and each column of $\boldsymbol{X}$ represents the set of features from a given observation; then, after a successful factorization, the columns of $\boldsymbol{B}$ can be thought of as recurrent signatures, forming a basis for the observations, and the rows of $\boldsymbol{G}$ as coefficients denoting the activation of each corresponding signature [9]. The presence of a given signature could then be related to some hidden factor of the underlying data; for instance, in a spectrogram factorization, an extracted signature can be linked to a particular emission from a musical instrument.

Often, however, the problem is so underdetermined that simply relying on the fit yielded by the minimization of $\mathcal{L}$ is not satisfactory. A common solution is then to restrict the model, imposing extra constraints on the obtained factors, such as sparsity, smoothness, or orthogonality [16, 17, 18].

One particular limitation of the simple NMF model is the fact that the extracted patterns are static, that is, the shape of the estimate provided by each component $\boldsymbol{b}_k$ is equal across all observations. Taking temporal continuity into consideration, such that the sequence of observations (columns of $\boldsymbol{X}$) are in fact a time sequence, a more powerful model can be designed, allowing the patterns to be longer in length than a single observation. The deconvolutive-NMF is a generalization that provides such features by extending the basis vectors $\boldsymbol{b}_k$ to matrices $\boldsymbol{B}_k \in \mathbb{R}_+^{I \times T}$, for some user-chosen T; the corresponding activation vector $\boldsymbol{g}_k$ is then time-shifted by t and applied to each subcomponent $\boldsymbol{b}_{kt}$:

$$\boldsymbol{X} \approx \sum_{k=1}^{K} \sum_{t=1}^{T} \boldsymbol{b}_{kt} \overrightarrow{\boldsymbol{g}_k^{\mathrm{T}}}^{t-1}; \tag{4}$$

the dataflow for the deconvolutive NMF is illustrated in Fig. 2.

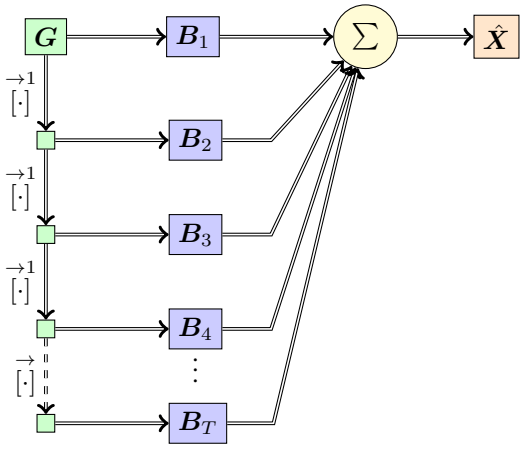

Fig. 2. Deconvolutive NMF model generating the approximation $\hat{\boldsymbol{X}}$. Here, the overall basis tensor $\boldsymbol{\mathcal{B}}$ is sliced across the deconvolutive mode, indexed by t.

The right-shift operator $\overset{t}{\rightarrow}[\cdot]$ is such that for a given vector

$$\overset{1}{\rightarrow}[g_1, g_2, \ldots, g_L] = [0, g_1, \ldots, g_{L-1}];$$

for $t = 0$, the operator reduces to identity, and with $t > 1$, it is equivalent to t successive applications of the operator with unit shift. The left-shift operator is defined analogously, and the extension to matrices and tensors follows naturally, shifting over some specified dimension.

The main idea behind deconvolutive-NMF in musical analysis is that longer musical note profiles can be efficiently modeled by the framework, leading to an increase in separation accuracy.

IV. Signal representation

We now describe the steps toward a phase-aware non-negative multichannel representation, aimed at the decomposition in non-negative factors. The first thing to do is to extend the basic data matrix $\boldsymbol{X}$, introduced in Section III, to the multichannel case. One way to do this is by associating each entry of this data matrix with an $M \times M$ matrix modeling relations between pairs of sensors; but then we also have to extend the notion of non-negativeness. When dealing with real numbers, being non-negative means belonging to the convex cone[1] $\mathbb{R}_+$. Positive-semidefiniteness is the matrix counterpart of the scalar non-negativeness property; indeed, positive-semidefiniteness is preserved under conical combination of positive-semidefinite (PSD) matrices.[2]

Translating the relationship of Eq. (1) to the short-time Fourier transform (STFT) domain, each (complex) frequency-time point measurement is

$$x_{ilm} = \sum_{q=1}^{Q} h_{iqm} s_{ilq}, \tag{5}$$

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 frequency-time point (i, l).

In order to represent the overall measurements as Hermitian PSD matrices, Sawada et al. [11] take the outer product of the vector $\boldsymbol{x}_{il}$—formed by the measurements across all sensors at a single frequency-time point—with itself, obtaining the matrices

$$\boldsymbol{P}_{il} = \boldsymbol{x}_{il}\boldsymbol{x}_{il}^{\mathrm{H}} = \sum_{q=1}^{Q}\sum_{q'=1}^{Q} \boldsymbol{h}_{iq}\boldsymbol{h}_{iq'}^{\mathrm{H}} s_{ilq}\bar{s}_{ilq'}, \tag{6}$$

where $\boldsymbol{v}^{\mathrm{H}}$ is the Hermitian of $\boldsymbol{v}$ and $\bar{z}$ is the complex conjugate of z. Additionally, assuming that different sources q and q' are orthogonal, and then transposing the correlation property to the STFT coefficients, a useful approximation can be obtained by neglecting the crossed terms:

$$\boldsymbol{P}_{il} \approx \sum_{q=1}^{Q} \boldsymbol{h}_{iq}\boldsymbol{h}_{iq}^{\mathrm{H}} |s_{ilq}|^2. \tag{7}$$

Lastly, in order to factor a magnitude instead of a power spectrum, the STFT coefficients are mapped through a magnitude square-root function applied elementwise, as proposed in [11]: $\phi(z) = \frac{z}{\sqrt{|z|}}$. In this new context, equation (6) is replaced by the corresponding expression

$$\boldsymbol{X}_{il} = \phi(\boldsymbol{x}_{il})\phi(\boldsymbol{x}_{il})^{\mathrm{H}} \approx \sum_{q=1}^{Q} \boldsymbol{\eta}_{iq}\boldsymbol{\eta}_{iq}^{\mathrm{H}} |s_{ilq}|, \tag{8}$$

where $\boldsymbol{\eta}_{iq}$ is a factor similar to $\boldsymbol{h}_{iq}$, allowing to define $\boldsymbol{H}_{iq} = \boldsymbol{\eta}_{iq}\boldsymbol{\eta}_{iq}^{\mathrm{H}}$ 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 each channel pair. By the outer product construction, $\boldsymbol{H}_{iq}$ preserves phase information without actually modeling the absolute phase of the measurements. Expression (8) then motivates a joint factorization of the sources' magnitude spectra and spatial-property matrices.

V. Factorization model

We now describe the steps toward the factorization of the multichannel model in (8) into non-negative factors. The standard CNMF [11] simply attaches to each NMF component some phase information in the form of a set of PSD matrices, as will be described in Subsection V-A. A limitation of this approach is that no structure is imposed on the family of PSD matrices associated with each component; we intend to separate sources based on spatial cues, and one could expect some form of coherence in terms of how the spatial information is encoded within a single component, as shown in [12]. In Subsection V-B, prior knowledge about the sensor array geometry is incorporated into the process in the form of beamforming kernels, from which the spatial matrices associated with each component are derived from. Finally, in Subsection V-C, the complete factorization model is glued together, defining some useful additional constraints, and in Subsection V-D the process of recovering the source image spectrograms from the CNMF parameters is described.

The main idea of factorization model 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 [11], 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 phase-magnitude model for the measured data points can be written as

$$\boldsymbol{X}_{il} \approx \sum_{k=1}^{K} \boldsymbol{H}_{ik}\check{s}_{ilk}, \tag{9}$$

where $\boldsymbol{H}_{ik}$ encodes spatial properties for a component and $\check{s}_{ilk}$ is a magnitude estimate that shall be computed through

[1] A subset $\mathcal{C}$ of a given vector space is a cone if $\alpha\boldsymbol{x} \in \mathcal{C}$ for any $\boldsymbol{x} \in \mathcal{C}$ and any $\alpha \geq 0$. It is a convex cone when it is closed for conical combinations, i.e., when $\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 \in \mathcal{C}$ for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{C}$ and any $\alpha_1, \alpha_2 \geq 0$.

[2] The space of PSD matrices, $\mathbb{S}_+$, is itself a cone, that is, any element $\boldsymbol{S} \in \mathbb{S}_+$ can be written as a conical combination of some basis matrices.

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 $\boldsymbol{X} \in \mathbb{R}_+^{I\times L}$ 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, $K \ll \mathrm{Rank}(\boldsymbol{X})$. 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 as presented in Section III. Using the deconvolutive model, the estimate of the instantaneous magnitudes due to the k^{th} component is

$$\check{s}_{ilk} = \sum_{t=1}^{T} b_{itk}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l. \tag{10}$$

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

$$\hat{\boldsymbol{X}}_{il} = \sum_{k=1}^{K} \boldsymbol{H}_{ik} \sum_{t=1}^{T} b_{itk}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l, \tag{11}$$

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

B. Spatial covariance model

We apply the direction-of-arrival (DoA) based factorization method introduced by Nikunen and Virtanen [12] 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. 3: for 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 relative time delay between sensor acquisitions can be defined in terms of the sensor array geometry, wave propagation velocity, and incidence direction. In fact, the difference in propagation length can be calculated as 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

$$\tau_{mm'}(\boldsymbol{k}_o) = \frac{\langle \boldsymbol{p}_{m'} - \boldsymbol{p}_m, \boldsymbol{k}_o \rangle}{c}, \tag{12}$$

where c denotes the wave propagation velocity.

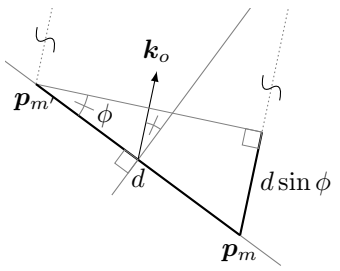

Fig. 3. TDoA as function of array geometry and wave incidence direction.

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

$$\theta_{mm'}(f_i, \boldsymbol{k}_o) = -2\pi f_i \tau_{mm'}(\boldsymbol{k}_o). \tag{13}$$

The idea for modeling $\boldsymbol{\mathcal{H}}$ 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 described in terms of the DoA kernels as conic combinations

$$\boldsymbol{H}_{ik} = \sum_{o=1}^{O} z_{ko}\boldsymbol{W}_{io}, \tag{14}$$

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

$$\hat{\boldsymbol{X}}_{il} = \sum_{k=1}^{K}\sum_{o=1}^{O} z_{ko}\boldsymbol{W}_{io} \sum_{t=1}^{T} b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l\,. \tag{15}$$

This factorization has a possible scaling ambiguity, so additional constraints are introduced, namely $\sum_o z_{ko}^2 = 1$, $\sum_l g_{kl}^2 = 1$, and $\|\boldsymbol{W}_{io}\|_{\mathrm{F}} = 1$, where $\|\boldsymbol{M}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\boldsymbol{M}^{\mathrm{H}}\boldsymbol{M})}$ denotes the Frobenius norm of a matrix $\boldsymbol{M}$. In order to find the best estimates for the parameters, a statistical model can be employed, and an estimation framework can be built—see Section VI.

D. Source reconstruction

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

$$\boldsymbol{y}_{ilq} = \boldsymbol{x}_{il} \frac{\sum_{k,o} \beta_{qk} z_{ko} \sum_t b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}{\sum_{q,k,o} \beta_{qk} z_{ko} \sum_t b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}, \tag{16}$$

where β_{qk} are learned membership coefficients relating a given component k to 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 by a mask of ratios of estimated spectral magnitudes, preserving the original phase. Finally, the discrete-time version of $\hat{y}_{qm}(t)$ can be retrieved from the inverse-STFT of the coefficients.

VI. Parameter estimation

Following previous works [11, 12], the generative model adopted for the entries of the data matrices $\boldsymbol{X}_{il}$ is that of independent complex Gaussian variables with unit variance, which allows to recast the likelihood estimate as a squared error minimization problem. In fact, the likelihood function for the parameters, considering the overall measurements, can be written as

$$\mathcal{L}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \propto \prod_{i=1}^{I} \prod_{l=1}^{L} \exp\left(-\|\boldsymbol{X}_{il} - \hat{\boldsymbol{X}}_{il}\|_{\mathrm{F}}^2\right), \tag{17}$$

leading to the alternative cost function

$$\ell(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) = \sum_{i=1}^{I} \sum_{l=1}^{L} \|\boldsymbol{X}_{il} - \hat{\boldsymbol{X}}_{il}\|_{\mathrm{F}}^2, \tag{18}$$

related to the colog-likelihood. It can be useful to embed some prior on the parameters. This knowledge can be directly related to a regularization factor, steering the algorithm toward 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[3]

$$\mathcal{L}_{\mathrm{R}} \propto \exp\left(-2\alpha_{\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_1\right) \mathcal{L}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \tag{19}$$

and a regularized cost function

$$\ell_{\mathrm{R}}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) = 2\alpha_{\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_1 + \sum_{i=1}^{I} \sum_{l=1}^{L} \|\boldsymbol{X}_{il} - \hat{\boldsymbol{X}}_{il}\|_{\mathrm{F}}^2, \tag{20}$$

where the tensor ℓ_1-norm is defined as $\sum_{i,k,t} |b_{ikt}|$. This is inspired by a LASSO [19] regression, aiming to enforce the selectivity of signatures $\boldsymbol{B}_k$ and, consequently, produce a sparser representation, suitable to pitched audio signals.

[3]We extract a factor 2 from $\alpha_{\boldsymbol{\mathcal{B}}}$ for convenience.

A. Minimization procedure

To obtain the maximum a posteriori estimate, one must minimize the regularized cost (20). The convoluted interdependence among the parameters $\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}$ in this cost function hinders its direct (gradient-based) minimization, calling for alternative procedures. A possible alternative would be to minimize an auxiliary function, related to the actual cost function, that depends on an extra set of parameters while being, in some sense, smoother. The individual gradients w.r.t. the CNMF parameters of the auxiliary function can be straightforwardly derived, and the extra set of parameters can be chosen in a way that explicit computation is avoided. In this subsection, we shall describe a convenient auxiliary function, along with the specific choice of additional parameters as well as the update rules for the non-negative tensors. At last, the update for the kernel matrices, which requires additional projection and normalization steps, is described.

While (20) 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 [20] may be employed with good results. We consider an auxiliary function to (20), namely:

$$\ell_{\mathrm{R}}^{+} = 2\alpha_{\boldsymbol{\mathcal{B}}} \|\boldsymbol{\mathcal{B}}\|_1 + \sum_{i,l,k,o,t} \frac{1}{r_{ilkot}} \|\boldsymbol{S}_{ilkot} - z_{ko} \boldsymbol{\mathcal{W}}_{io} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l\|_{\mathrm{F}}^2, \tag{21}$$

where r_{ilkot} are any positive variables satisfying $\sum_{k,o,t} r_{ilkot} = 1$, and $\boldsymbol{S}_{ilkot}$ are Hermitian matrices satisfying $\sum_{k,o,t} \boldsymbol{S}_{ilkot} = \boldsymbol{X}_{il}$, for which we derive the following proposition:

Proposition 1. *For all $\boldsymbol{S}_{ilkot} \in \mathbb{S}$ and $r_{ilkot} \in \mathbb{R}_+$ such that $\sum_{k,o,t} \boldsymbol{S}_{ilkot} = \boldsymbol{X}_{il}$ and $\sum_{k,o,t} r_{ilkot} = 1$, then*

$$\ell_{\mathrm{R}}^{+}(\boldsymbol{\mathcal{S}}, \boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \geq \ell_{\mathrm{R}}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \quad \textit{and} \tag{22}$$

$$\min_{\boldsymbol{\mathcal{S}}} \ell_{\mathrm{R}}^{+}(\boldsymbol{\mathcal{S}}, \boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) = \ell_{\mathrm{R}}(\boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}). \tag{23}$$

Furthermore,

$$\begin{aligned} \boldsymbol{S}_{ilkot}^{*} &\triangleq \arg\min_{\boldsymbol{S}_{ilkot}} \ell_{\mathrm{R}}^{+}(\boldsymbol{\mathcal{S}}, \boldsymbol{Z}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{B}}, \boldsymbol{G}) \\ &= z_{ko} \boldsymbol{\mathcal{W}}_{io} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l - r_{ilkot} \boldsymbol{E}_{il}. \end{aligned} \tag{24}$$

Proof. See Appendix A. □

Through the majorizer conditions (22) and (23), the individual minimization of (21) across the CNMF variables with $\boldsymbol{\mathcal{S}}$ set as the optimal $\boldsymbol{\mathcal{S}}^*$ is guaranteed to be non-increasing. With the auxiliary definition

$$\hat{x}_{il} = \sum_{k,o,t} z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l, \tag{25}$$

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

$$r_{ilkot} = \frac{z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}{\hat{x}_{il}}. \tag{26}$$

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

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

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 (see Appendix B):

$$z_{ko} \longleftarrow z_{ko}\left[\frac{\sum_{i,l,t}\left(\hat{x}_{il} + \operatorname{tr}\left(\boldsymbol{E}_{il}\boldsymbol{W}_{io}\right)\right) b_{ikt}[\overset{t-1}{\rightarrow}\boldsymbol{g}_k]_l}{\sum_{i,l,t}\hat{x}_{il}b_{ikt}[\overset{t-1}{\rightarrow}\boldsymbol{g}_k]_l}\right], \tag{27}$$

$$b_{ikt} \longleftarrow b_{ikt}\left[\frac{\sum_{l,o}\left(\hat{x}_{il} + \operatorname{tr}\left(\boldsymbol{E}_{il}\boldsymbol{W}_{io}\right)\right) z_{ko}[\overset{t-1}{\rightarrow}\boldsymbol{g}_k]_l}{\alpha_{\boldsymbol{\mathcal{B}}} + \sum_{l,o}\hat{x}_{il}z_{ko}[\overset{t-1}{\rightarrow}\boldsymbol{g}_k]_l}\right], \tag{28}$$

$$g_{kl} \longleftarrow g_{kl}\left[\frac{\sum_{i,o,t}([\overset{t-1}{\leftarrow}\hat{\boldsymbol{x}}_i]_l + \operatorname{tr}([\overset{t-1}{\leftarrow}\boldsymbol{\mathcal{E}}_i]_l\boldsymbol{W}_{io}))z_{ko}b_{ikt}}{\sum_{i,o,t}[\overset{t-1}{\leftarrow}\hat{\boldsymbol{x}}_i]_l z_{ko}b_{ikt}}\right]. \tag{29}$$

The update process for the kernel matrices is slightly different, since only magnitude optimizations 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 (see Appendix C)

$$\hat{\boldsymbol{W}}_{io} \longleftarrow \frac{\sum_{l,k,t} z_{ko}b_{ikt}[\overset{t-1}{\rightarrow}\boldsymbol{g}_k]_l[\hat{x}_{il}\boldsymbol{W}_{io} + \boldsymbol{E}_{il}]}{\sum_{l,k,t}\hat{x}_{il}z_{ko}b_{ikt}[\overset{t-1}{\rightarrow}\boldsymbol{g}_k]_l}; \tag{30}$$

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

$$\boldsymbol{V}_{io}\boldsymbol{\Lambda}_{io}\boldsymbol{V}^{\mathrm{H}}_{io} \longleftarrow \hat{\boldsymbol{W}}_{io} \tag{31}$$

$$\hat{\boldsymbol{W}}^{+}_{io} \longleftarrow \boldsymbol{V}_{io}\boldsymbol{\Lambda}^{+}_{io}\boldsymbol{V}^{\mathrm{H}}_{io}; \tag{32}$$

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

$$\boldsymbol{W}_{io} \longleftarrow \operatorname{abs}(\hat{\boldsymbol{W}}^{+}_{io}) \odot \operatorname{sign}(\boldsymbol{W}_{io}), \tag{33}$$

where $\operatorname{abs}(\cdot)$ and $\operatorname{sign}(\cdot)$ both 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:

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

VII. Numerical results

The program developed to assess the accuracy of the proposals was coded in Python, using TensorFlow v1.14, and executed on an Intel Xeon Gold 5120. We attempt the separation of two sound sources positioned 90 degrees apart, from four mixtures synchronously captured by omnidirectional microphones placed approximately 8 cm apart from each other, in the form of a tetrahedron. The audio tracks are two musical samples about 20 s long, sampled at 22.05 kHz. A closed room of dimensions 5 m × 4 m × 3 m (length, width, height) with $\mathrm{RT}_{60} \approx 450$ ms was simulated using CATT-Acoustic v9.0c [21], with the microphone array at the center and the sources on the horizontal plane 1.5 m away from it, as depicted in Fig. 4. The DoA vectors were randomly sampled from the unit sphere in a way to approximately maximize the cosine distance between the closest vectors, with $O = 110$ directions. An illustration of the DoA coordinates is shown in Fig. 5.

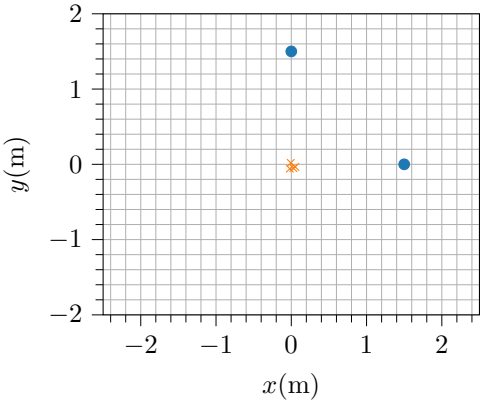

Fig. 4. Top-down view of the source-sensors arrangement inside the room. The array is indicated by orange cross markers, and the two sources by blue dots. The reference axes are aligned with the walls of the room.

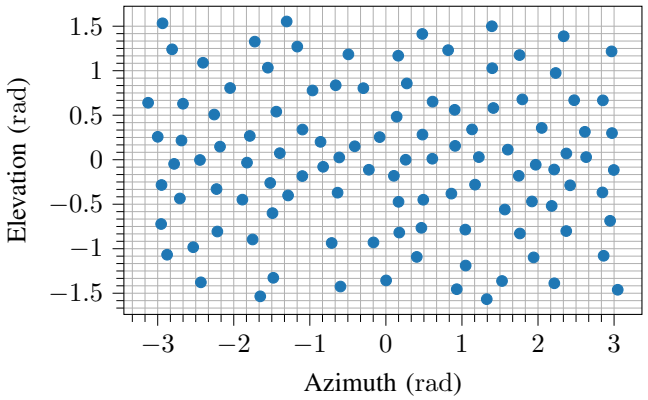

Fig. 5. Coordinates of the DoA samples on the sphere.

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 $L = 978$ frames. We considered three scenarios: a baseline run, with $(K = 60, T = 1, \alpha_{\boldsymbol{\mathcal{B}}} = 0)$, equivalent to [12]; a deconvolutive run, with $(K = 60, T = 5, \alpha_{\boldsymbol{\mathcal{B}}} = 0)$ testing the proposed extension; and a sparsity-promoting run,

with $(K = 60, T = 5, \alpha_{\boldsymbol{\mathcal{B}}} = 0.5)$. The algorithm was run for 500 iterations in all cases.

We performed weighted k-means on the vectors $\boldsymbol{z}_k$, with weights corresponding to the component energy $\|\boldsymbol{B}_k\|_{\mathrm{F}}^2$. The source-component membership coefficients β_{qk} were set to 1 or 0 based on the obtained clustering.

First, we analyze the directional sensitivity of the method. With k-means clustering, it is possible to retrieve the centroids corresponding to the spatial signatures for each source; based on the spherical coordinates of the DoA vectors and the obtained centroid coefficients, a bivariate spherical spline interpolation was used to create a visualization of the spatial properties for each source. The results are depicted in Figs. 6 and 7. The true positions are depicted with blue dots.

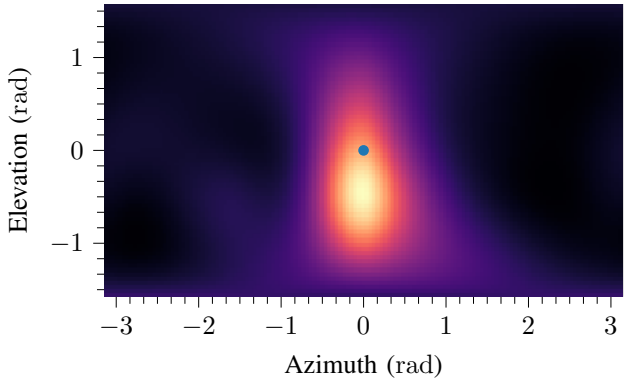

Fig. 6. Visualization of spatial signature for the first source. A main lobe is clearly discernible, aligned with the source's true direction (blue dot).

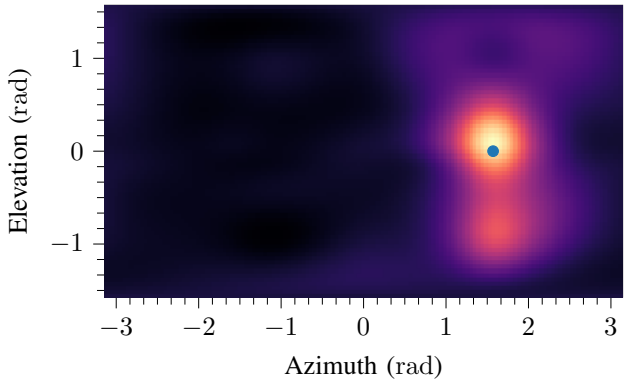

Fig. 7. Visualization of spatial signature for the second source. A main lobe is clearly discernible, closely matching the source's true direction (blue dot).

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 [22, 23] was measured using the `mir_eval` suite [24, 25]. From the suite's provided image evaluation function, we obtain the source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) for each source, and the results are depicted in Fig. 8. SDR encompasses several types of distortions into a single metric, being a robust form of evaluation for BSS techniques; SIR, on the other hand measures only the overall crosstalk between estimated sources, being another useful metric for separation evaluation.

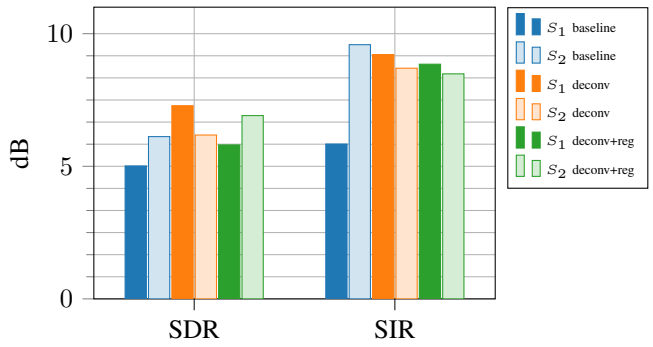

Fig. 8. Separation metrics for the three tests. Both deconvolutive and regularized tests obtained superior averages than the baseline, while the regularized factorization was slightly worse than the non-regtest.

Due to the stochastic nature of the fit, an unbalance between sources can be noted in all scenarios, but a slight advantage can be noted in the deconvolutive case. The regularization slightly degraded the overall measures in comparison to the unregularized deconvolutive test, although they are still superior (on average) to those of the baseline case.

We would like to investigate the sparsity-inducing properties on the extracted signatures $\boldsymbol{\mathcal{B}}$. While the factors obtained by fitting the data using the Euclidean cost are often sparse, they are not the sparsest possible representation. Due to the nature of the range of the values assumed by b_{ikt}, we ploted the histogram of $\log_{10}(b_{ikt})$ using 50 bins as shown in Fig. 9. A considerable increase in frequency of the lowest bin can be observed: the count on the lowest bin for the unregularized run was 84991 hits, while the regularized test had 145982 hits, a 71% increase. It is then clear that the regularizing component of $\mathcal{L}_{\mathrm{R}}$ steered the factor search toward a solution with higher selectivity.

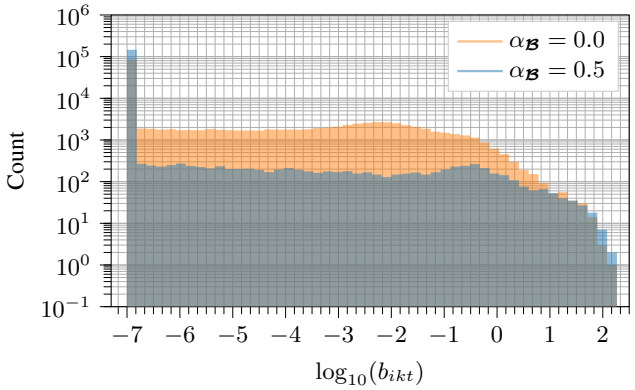

Fig. 9. Histogram of $\boldsymbol{\mathcal{B}}$ coefficients. Due to numerical issues, coefficients lower than some $\epsilon = 10^{-7}$ were clipped to ϵ.

In general terms, one desires sparse signatures if the data have 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, the imposed regularization is expected to enhance the precision of the algorithm for tonal

emissions, while some degradation can occur in percussive or other atonal emissions. In our test scenarios, the regularization did result in some slight degradation in separation, suggesting that the tonal assumption is not strictly valid—which is not surprising in real-life composed signals. Nevertheless, ℓ_1 penalty parameters are ubiquitous in regular NMF algorithms [26, 27], so that the added model flexibility is welcome.

VIII. Conclusions

We proposed an extended version [1] of the CNMF algorithm [11], leveraging the efficient representation from the deconvolutive NMF model. We also provided the user with control over the distribution of the extracted signatures through regularization, which steers the method toward 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.

Although testing was not exhaustive, due to the large number of user-chosen parameters and different possible configurations, our simulations corroborated the method's capabilities in the separation task.

Future works include the application of these ideas using different generative models for the matrices $\boldsymbol{X}_{il}$, as the Gaussian assumption and the associated Frobenius norm method are not entirely adequate for distance measures on the PSD cone; however, the fact that, by construction, $\boldsymbol{X}_{il}$ is rank-1 and lives on the boundary of $\mathbb{S}_+$ imposes challenges on the application of traditional measures, such as $\log\det$ divergences [28], which are only defined in the interior of the cone. Also, application of other NMF frameworks, such as multi-layer NMF, could improve the accuracy and consistency of the method. Proper initialization is also a critical point in traditional NMF methods [29], and the design of a smart initialization routine is likely to have a considerable effect on convergence time and accuracy.

Appendix A
Proof of Proposition 1

First, the Lagrangian function for Eq. (21) w.r.t. $\boldsymbol{S}_{ilkot}$ restriction is computed:

$$\psi_{il} = \sum_{k,o,t} \frac{\|\boldsymbol{S}_{ilkot} - z_{ko}\boldsymbol{W}_{io}b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l\|_{\mathrm{F}}^2}{r_{ilkot}} - \operatorname{tr}(\boldsymbol{M}_{il}(\boldsymbol{X}_{il} - \sum_{k,o,t}\boldsymbol{S}_{ilkot})^{\mathrm{H}}), \quad (34)$$

where the regularization term was omitted for brevity. The partial derivative w.r.t. $\boldsymbol{S}_{ilkot}$is given by

$$\frac{\partial\psi_{il}}{\partial\boldsymbol{S}_{ilkot}} = \frac{2(\boldsymbol{S}_{ilkot} - z_{ko}\boldsymbol{W}_{io}b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l)}{r_{ilkot}} - \boldsymbol{M}_{il}. \quad (35)$$

Summing over k, o, and t after equating to $\mathbf{0}$ results in

$$\begin{aligned}\sum_{k,o,t} r_{ilkot}\boldsymbol{M}_{il} &= 2(\sum_{k,o,t}\boldsymbol{S}_{ilkot} - \sum_{k,o,t} z_{ko}\boldsymbol{W}_{io}b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l) \\ \boldsymbol{M}_{il} &= 2(\boldsymbol{X}_{il} - \hat{\boldsymbol{X}}_{il}) = 2\boldsymbol{E}_{il}.\end{aligned} \quad (36)$$

Finally, backsolving Eq. (35) for $\boldsymbol{S}_{ilkot}$ yields

$$\boldsymbol{S}^*_{ilkot} = z_{ko}\boldsymbol{W}_{io}b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l - r_{ilkot}\boldsymbol{E}_{il}; \quad (37)$$

replacing $\boldsymbol{S}^*_{ilkot}$ in Eq. (21) leads directly to the equality in Eq. (23), and the fact that Eq. (21) is quadratic in $\boldsymbol{S}_{ilkot}$, with a single global minimum, is sufficient to satisfy inequation (22).

Appendix B
Derivation of Equations (27), (28), and (29)

The partial derivatives of ℓ_R^+ w.r.t. the non-negative parameters are:

$$\frac{\partial\ell_R^+}{\partial z_{ko}} = \sum_{i,l,t}\frac{-2}{r_{ilkot}}\left[-b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l \operatorname{tr}(\boldsymbol{S}_{ilkot}\boldsymbol{W}_{io}) + z_{ko}b_{ikt}^2[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l^2\right], \quad (38)$$

$$\frac{\partial\ell_R^+}{\partial b_{ikt}} = 2\alpha_{\boldsymbol{\mathcal{B}}} + \sum_{l,o}\frac{-2}{r_{ilkot}}\left[-z_{ko}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l \operatorname{tr}(\boldsymbol{S}_{ilkot}\boldsymbol{W}_{io}) + z_{ko}^2 b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l^2\right], \quad (39)$$

$$\frac{\partial\ell_R^+}{\partial g_{kl}} = \sum_{i,o,t}\frac{-2}{[\overleftarrow{\boldsymbol{r}_{ikot}}^{t-1}]_l}\left[-z_{ko}b_{ikt}\operatorname{tr}\left([\overleftarrow{\boldsymbol{\mathcal{S}}_{ikot}}^{t-1}]_l\boldsymbol{W}_{io}\right) + z_{ko}^2 b_{ikt}^2 g_{kl}\right], \quad (40)$$

(note that the left-shift operator appears in (40)), where the fact that $\operatorname{tr}(\boldsymbol{W}_{io}\boldsymbol{W}_{io}^{\mathrm{H}}) = 1$ was used.

Applying the definition of r_{ilkot} in terms of z_{ko}, b_{ikt}, g_{kl}, and $\hat{x}_{il}$, expressions of type $\frac{b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}{r_{ilkot}}$ are replaced with $\frac{\hat{x}_{il}}{z_{ko}}$, and those of type $\frac{z_{ko}b_{ikt}^2[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l^2}{r_{ilkot}}$ with $\hat{x}_{il}b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l$; solving for the respective variables yields:

$$z^*_{ko} = \frac{\sum_{i,l,t}\hat{x}_{il}\operatorname{tr}(\boldsymbol{S}_{ilkot}\boldsymbol{W}_{io})}{\sum_{i,l,t}\hat{x}_{il}b_{ikt}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}, \quad (41)$$

$$b^*_{ikt} = \frac{\sum_{l,o}\hat{x}_{il}\operatorname{tr}(\boldsymbol{S}_{ilkot}\boldsymbol{W}_{io})}{\alpha_{\boldsymbol{\mathcal{B}}} + \sum_{l,o}\hat{x}_{il}z_{ko}[\overrightarrow{\boldsymbol{g}_k}^{t-1}]_l}, \quad (42)$$

$$g^*_{kl} = \frac{\sum_{i,o,t}[\overleftarrow{\hat{\boldsymbol{x}}_i}^{t-1}]_l\operatorname{tr}\left([\overleftarrow{\boldsymbol{\mathcal{S}}_{ikot}}^{t-1}]_l\boldsymbol{W}_{io}\right)}{\sum_{i,o,t}[\overleftarrow{\hat{\boldsymbol{x}}_i}^{t-1}]_l z_{ko}b_{ikt}}. \quad (43)$$

Finally, replacing $\boldsymbol{\mathcal{S}}$ with $\boldsymbol{\mathcal{S}}^*$ leads to the multiplicative updates.

APPENDIX C
DERIVATION OF EQUATION (30)

The gradient of ℓ_R^+ w.r.t. $\boldsymbol{W}_{io}$ is

$$\frac{\partial \ell_R^+}{\partial \boldsymbol{W}_{io}} = \sum_{l,k,t} \frac{2}{r_{ilkot}} \left[-z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l \boldsymbol{S}_{ilkot} + z_{ko}^2 \boldsymbol{W}_{io} b_{ikt}^2 [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l^2 \right]. \quad (44)$$

Once more, applying the choice of r_{ilkot}, and solving for $\frac{\partial \ell_R^+}{\partial \boldsymbol{W}_{io}} = \mathbf{0}$ yields

$$\hat{\boldsymbol{W}}_{io} = \frac{\sum_{l,k,t} \hat{x}_{il} \boldsymbol{S}_{ilkot}}{\sum_{l,k,t} \hat{x}_{il} z_{ko} b_{ikt} [\overrightarrow{\boldsymbol{g}_k^{t-1}}]_l}. \quad (45)$$

Replacing $\boldsymbol{\mathcal{S}}$ with $\boldsymbol{\mathcal{S}}^*$ leads to the multiplicative update.

REFERENCES

[1] T. L. B. Dias, L. W. P. Biscainho, and W. A. Martins, "Time-deconvolutive CNMF for multichannel blind source separation," in *Anais do XXXVII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*. Petrópolis, Brazil: SBrT, Sep. 2019, pp. 1–5, doi: 10.14209/sbrt.2019.1570557364.

[2] M. Pal, R. Roy, J. Basu, and M. S. Bepari, "Blind source separation: A review and analysis," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. Gurgaon, India: IEEE, Nov. 2013, pp. 1–5, doi: 10.1109/ICSDA.2013.6709849.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, "What is independent component analysis?" in *Independent Component Analysis*. John Wiley & Sons, Ltd, 2002, ch. 7, pp. 145–164, doi: 10.1002/0471221317.ch7.

[4] ——, "ICA by maximization of nongaussianity," in *Independent Component Analysis*. John Wiley & Sons, Ltd, 2002, ch. 8, pp. 165–202, doi: 10.1002/0471221317.ch8.

[5] ——, "ICA by maximum likelihood estimation," in *Independent Component Analysis*. John Wiley & Sons, Ltd, 2002, ch. 9, pp. 203–219, doi: 10.1002/0471221317.ch9.

[6] 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, Dec. 2008, doi: 10.1016/j.clinph.2008.09.007.

[7] 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)*. Limassol, Cyprus: IEEE, Mar. 2010, pp. 1–6, doi: 10.1109/ISCCSP.2010.5463317.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.

[9] 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, doi: 10.5555/2700548.

[10] 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: IEEE, Oct. 2003, pp. 177–180, doi: 10.1109/ASPAA.2003.1285860.

[11] 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: IEEE, Oct. 2011, pp. 153–156, doi: 10.1109/ASPAA.2011.6082275.

[12] 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, Mar. 2014, doi: 10.1109/TASLP.2014.2303576.

[13] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 6677–6681, doi: 10.1109/ICASSP.2014.6854892.

[14] 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. Granada, Spain: Springer, Sep. 2004, pp. 494–499, doi: 10.1007/978-3-540-30110-3_63.

[15] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, "Introduction – problem statements and models," in *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, Ltd, 2009, ch. 1, pp. 1–80, doi: 10.1002/9780470747278.ch1.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, doi: 10.5555/1162264.

[17] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, Mar. 2010, doi: 10.1109/TASL.2010.2041381.

[18] J. Yoo and S. Choi, "Nonnegative matrix factorization with orthogonality constraints," *Journal of Computer Science and Engineering*, vol. 4, no. 2, pp. 97–109, Jun. 2010, doi: 10.5626/JCSE.2010.4.2.097.

[19] 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, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[20] 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, doi: 10.1007/978-3-642-46808-7_28.

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

[22] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006, doi: 10.1109/TSA.2005.858005.

[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 *Independent Component Analysis and Signal Separation*, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds. London, UK: Springer, Sep. 2007, pp. 552–559, doi: 10.1007/978-3-540-74494-8_69.

[24] 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*. Taipei, Taiwan: ISMIR, Oct. 2014, pp. 367–372, doi: 10.1.1.722.1267.

[25] C. Raffel, "mir_eval documentation," http://craffel.github.io/mir_eval/, [Online; accessed February 20, 2019].

[26] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, "Multiplicative iterative algorithms for NMF with sparsity constraints," in *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, Ltd, 2009, ch. 3, pp. 131–202, doi: 10.1002/9780470747278.ch3.

[27] ——, "Alternating least squares and related algorithms for NMF and SCA problems," in *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, Ltd, 2009, ch. 4, pp. 203–266, doi: 10.1002/9780470747278.ch4.

[28] A. Cichocki, S. Cruces, and S. ichi Amari, "Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences," *Entropy*, vol. 17, no. 5, pp. 2988–3034, May 2014, doi: 10.3390/e17052988.

[29] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, Apr. 2008, doi: 10.1016/j.patcog.2007.09.010.

Thadeu Luiz Barbosa Dias was born in Rio de Janeiro, Brazil, in 1994. B.Sc. degree in electronics and computer engineering from the Universidade Federal do Rio de Janeiro (POLI/UFRJ) in 2020, currently pursuing his M.Sc. degree in electrical engineering at COPPE (PEE/COPPE/UFRJ). Mr. Dias's research interests include digital signal processing, image processing, machine learning and information geometry.

Wallace Alves Martins received the Electronics Engineer degree from Universidade Federal do Rio de Janeiro (UFRJ, Brazil) in 2007, the M.Sc. and D.Sc. degrees in Electrical Engineering also from UFRJ in 2009 and 2011, respectively. He was a Research Visitor at University of Notre Dame (USA, 2008), at Université de Lille 1 (France, 2016), and at Universidad de Alcalá (Spain, 2018). From 2010 to 2013 he was an Associate Professor of Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ, Brazil). Since 2013 he has been with the Department of Electronics and Computer Engineering (DEL/Poli) and Electrical Engineering Program (PEE/COPPE) at UFRJ, where he is presently an Associate Professor (on leave). He is currently a Research Associate working with the Interdisciplinary Centre for Security, Reliability and Trust (SnT) at Université du Luxembourg. His research interests are in the fields of digital signal processing, especially adaptive signal processing and graph signal processing, as well as digital communications, with focus on equalization and precoding for wireless communications. Dr. Martins has authored more than 60 technical papers in refereed international journals and conferences, and 1 scientific book. Also, he received the Best Student Paper Award from EURASIP at EUSIPCO-2009, Glasgow, Scotland, and the 2011 Best Brazilian D.Sc. Dissertation Award from Capes. He is currently a member of IEEE (Institute of Electrical and Electronics Engineers) and SBrT (Brazilian Telecommunications Society).

Luiz Wagner Pereira Biscainho was born in Rio de Janeiro, Brazil, in 1962. He received the diploma of electronic engineer (magna cum laude) from the EE (now Poli) at Universidade Federal do Rio de Janeiro (UFRJ), Brazil, in 1985, and the M.Sc. and D.Sc. degrees in electrical engineering from the COPPE at UFRJ in 1990 and 2000, respectively. Having worked in the telecommunication industry between 1985 and 1993, Dr. Biscainho is now Associate Professor at DEL/Poli and PEE/COPPE, at UFRJ. His research area is digital signal processing, particularly audio processing. He is currently a member of IEEE (Institute of Electrical and Electronics Engineers), AES (Audio Engineering Society), SBrT (Brazilian Telecommunications Society), and SBC (Brazilian Computer Society).