Gaussian modeling-based multichannel audio source separation exploiting generic source spectral model

Thanh T. H. Duong, Ngoc Q. K. Duong *Senior Member, IEEE*, Phuong Cong Nguyen, and Cuong Quoc Nguyen

***Abstract*—As *blind* audio source separation has remained very challenging in real-world scenarios, some existing works, including ours, have investigated the use of a *weakly-informed* approach where generic source spectral models (GSSM) can be learned a priori based on nonnegative matrix factorization (NMF). Such approach was derived for single-channel audio mixtures and shown to be efficient in different settings. This paper proposes a multichannel source separation approach where the GSSM is combined with the source spatial covariance model within a unified Gaussian modeling framework. We present the generalized expectation-minimization (EM) algorithm for the parameter estimation. Especially, for guiding the estimation of the intermediate source variances in each EM iteration, we investigate the use of two criteria: (1) the estimated variances of each source are constrained by NMF, and (2) the total variances of all sources are constrained by NMF altogether. While the former can be seen as a source variance denoising step, the latter is viewed as an additional separation step applied to the source variance. We demonstrate the speech separation performance, together with its convergence and stability with respect to parameter setting, of the proposed approach using a benchmark dataset provided within the 2016 Signal Separation Evaluation Campaign.**

KEYWORDS

Multichannel audio source separation, local Gaussian model, nonnegative matrix factorization, generic spectral model, group sparsity constraint.

I. INTRODUCTION

Real-world recordings are often mixtures of several audio sources, which usually deteriorate the target one. Thus many practical applications such as speech enhancement, sound post-production, and robotics use audio source separation technique [1], [2] to extract individual sound sources from their mixture. However, despite numerous effort in the past decades, blind source separation performance in reverberant recording conditions is still far from perfect [3], [4]. To improve the separation performance, *informed* approaches have been proposed and emerged recently in the literature [5], [6]. Such approaches exploit side information about either the sources themselves or the mixing condition in order to guide the separation process. Examples of the investigated side information include deformed or hummed references of one (or more) source(s) in a given mixture [7], [8], text associated with spoken speeches [9], temporal annotation of the source activity along the mixtures [10], core associated with musical sources [11], [12], and motion associated with audio-visual objects in a video [13]. Following this trend, some recent works including ours have proposed to use a very abstract semantic information just about the types of audio sources existing in the mixture to guide the source separation. If one source in the mixture is known as "speech", then several speaker-independent speech examples can be used to create a universal speech model as presented in [14]; if several types of sound sources in the mixture are known (*e.g.,* birdsong, piano, waterfall), their audio examples found by internet search can be used to learn the corresponding universal sound class models as presented in [15]. Such universal models were shown to be effective in guiding the source separation algorithm and resulted in promising performance. Inspired by this idea, we have further investigated the use of generic speech and noise model for single-channel speech separation in [16] and shown its promising result in (a) the supervised case, where both speech GSSM and noise GSSM are learned during training phase, and (b) the semi-supervised case, where only the speech GSSM is pre-learned. Furthermore, we have proposed to combine the block sparsity constraint investigated in [14] with the component sparsity constraint presented in [17] in a common formulation in order to take into account the advantage of both of them [18].

It should be noted that the works cited above [9], [12], [16], [18] considered only a single channel case, where the mixtures are mono, and exploited non-negative matrix factorization (NMF) [19], [20] to model the spectral characteristics of audio sources. Some recent works have investigated the use of the deep neural networks (DNN) to model the source spectra, where basically the types of sources in the mixture also need to be known as a side information in order to collect training data. Such DNN-based approaches were shown to offer very promising results in single-channel speech and music separation [21]–[23], multichannel speech separation [24], [25]. However, they require a large amount of labeled data for training, which may not always be available and the training is usually computationally expensive.

When more recording channels are available thanks to the use of multiple microphones, a multichannel source separation algorithm should be considered as it allows to exploit important information about the spatial locations of audio sources. Such spatial information is reflected in the mixing process (usually with reverberation), and can be modeled by

Thanh T. H. Duong is with Hanoi University of Mining and Geology and International Research Institute MICA, Vietnam, e-mail: (duongthihienthanh@humg.edu.vn).

Ngoc Q. K. Duong is with Technicolor R&I, France, e-mail: (quang-khanh-ngoc.duong@technicolor.com).

Phuong Cong Nguyen is with MICA and Hanoi University of Science and Technology, Vietnam, e-mail: (phuong.nguyencong@hust.edu.vn).

Cuong Quoc Nguyen is with Hanoi University of Science and Technology, Vietnam, e-mail: (cuong.nguyenquoc@hust.edu.vn).

e.g., the interchannel time difference (ITD) and interchannel intensity difference (IID) [26]–[29], the rank-1 time-invariant mixing vector in the frequency domain when following the narrowband assumption [30]–[33], or the full-rank spatial covariance matrix in local Gaussian model (LGM) where the narrowband assumption is relaxed [34]–[36].

In this paper, we present an extension of the previous works [15], [16], [18] to the multichannel case where the NMF-based GSSM is combined with the full-rank spatial covariance model in a Gaussian modeling paradigm. Around this LGM, existing works have investigated several source spectral models such as Gaussian mixture model (GMM) [37], NMF as a linear model with nonnegativity constraints [36], [38], continuity model [39], kernel additive model [40], heavy-tailed distributions-based model [41], [42], and recently DNN [24]. Focusing on NMF in this study, our work is most closely related to [38] and [36] as both of them use NMF within the LGM to constrain the source spectra in each EM iteration. However, our work is different from [38] in the sense that we use the pre-trained GSSM, so that potentially the algorithm is less sensitive to the parameter initialization, and it does not suffer from the well-known permutation problem. Our work is also different from [36] as we exploit the mixed group sparsity constraint to guide the optimization, which allows the algorithm to automatically select the most representative spectral components in the GSSM. In addition, instead of constraining the variances of each source by NMF as done in [36], [38], we propose to constrain the total variances of all sources altogether by NMF and show that this novel optimization criterion offers better source separation performance. While part of the work was presented in [43], this paper provides more details regarding the algorithm derivation and the parameter settings. Furthermore, the source separation performance analysis and the comparison with existing approaches are extended.

The rest of the paper is organized as follows. We discuss the problem formulation and the background in Section II. We present the proposed GSSM-based multichannel source separation approach in Section III. In this section, we first present two ways of constructing the GSSM based on NMF. Then, to constrain the intermediate source variance estimates, two optimization criteria are introduced, which can be seen as either performing source variance denoising or source variance separation. The generalized EM algorithm is derived for the parameter estimation. We finally validate the effectiveness of the proposed approach in speech enhancement scenario using a benchmark dataset from the 2016 Signal Separation Evaluation Campaign (SiSEC 2016) in Section IV. For this purpose, we first analyze the convergence of the derived algorithm and investigate its sensitivity to the parameter settings in terms of source separation performance. We then show that the proposed algorithm outperforms most state-of-the-art methods in terms of the energy-based criteria.

II. PROBLEM FORMULATION AND MODELING

In this section, we review the formulation and the Gaussian modeling framework for multichannel audio source separation. Let us formulate the problem in a general setting, where J sources are observed by an array of I microphones. The contribution of each source, indexed by j, to the microphone array is denoted by a vector $\mathbf{c}_j(t) \in \mathbb{R}^{I\times 1}$ and the I-channel mixture signal is the sum of all source images as

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t). \tag{1}$$

The objective of source separation is to estimate the source images $\mathbf{c}_j(t)$ given $\mathbf{x}(t)$. As the considered algorithm operates in the frequency domain, we denote by $\mathbf{c}_j(n,f)$ and $\mathbf{x}(n,f)$ the complex-valued short-term Fourier transforms (STFT) of $\mathbf{c}_j(t)$ and $\mathbf{x}(t)$, respectively, where $n = 1,2,..,N$ is time frame index and $f = 1,2,...,F$ the frequency bin index. Equation (1) can be written in the frequency domain as

$$\mathbf{x}(n,f) = \sum_{j=1}^{J} \mathbf{c}_j(n,f). \tag{2}$$

A. Local Gaussian model

We consider the existing nonstationary LGM as it has been known to be robust in modeling reverberant mixing conditions and flexible in handling prior information [34], [37]. In this framework, $\mathbf{c}_j(n,f)$ is modeled as a zero-mean complex Gaussian random vector with covariance matrix $\mathbf{\Sigma}_j(n,f) = \mathbb{E}(\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f))$:

$$\mathbf{c}_j(n,f) \sim \mathcal{N}_c(\mathbf{0}, \mathbf{\Sigma}_j(n,f)), \tag{3}$$

where $\mathbf{0}$ is an $I \times 1$ vector of zeros and ^{H} indicates the conjugate transposition. Furthermore, the covariance matrix is factorized as

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f), \tag{4}$$

where $v_j(n,f)$ are scalar time-dependent *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(f)$ are time-independent $I \times I$ *spatial covariance matrices* encoding their spatial characteristics when sources and microphones are assumed to be static. Under the assumption that the source images are statistically independent, the mixture vector $\mathbf{x}(n,f)$ also follows a zero-mean multivariate complex Gaussian distribution with the covariance matrix computed as

$$\mathbf{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{5}$$

Assuming that the mixture STFT coefficients at all time-frequency (T-F) bins are independent, the likelihood of the set of observed mixture vectors $\mathbf{x} = \{\mathbf{x}(n,f)\}_{n,f}$ given the set of variance and spatial covariance parameters $\theta = \{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ is given by

$$P(\mathbf{x}|\theta) = \prod_{n,f} \frac{1}{\det\left(\pi \mathbf{\Sigma}_{\mathbf{x}}(n,f)\right)} e^{-\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f))}, \tag{6}$$

where det represents determinant of a matrix, tr() stands for matrix trace, and $\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f) = \mathbb{E}(\mathbf{x}(n,f)\mathbf{x}^H(n,f))$ is the empirical covariance matrix. It can be numerically computed

by local averaging over neighborhood of each T-F bin (n', f') as [36], [44]:

$$\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f) = \sum_{n',f'} w_{nf}^2(n',f')\mathbf{x}(n',f')\mathbf{x}^H(n',f'), \quad (7)$$

where w_{nf} is a bi-dimensional window specifying the shape of the neighborhood such that $\sum_{n',f'} w_{nf}^2(n',f') = 1$. We use Hanning window in our implementation. The quadratic T-F presentation as $\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)$ aims to improve the robustness of the parameter estimation as it exploits the observed data in several T-F points instead of a single one. The negative log-likelihood derived from (6) is

$$\mathcal{L}(\theta) = \sum_{n,f} \mathrm{tr}\big(\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)\big) + \log\det\big(\pi\mathbf{\Sigma}_{\mathbf{x}}(n,f)\big), \quad (8)$$

Under this model, once the parameters θ are estimated, the STFT coefficients of the source images are obtained in the minimum mean square error (MMSE) sense by multichannel Wiener filtering as

$$\hat{\mathbf{c}}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \quad (9)$$

Finally, the expected time-domain source images $\hat{\mathbf{c}}_j(t)$ are obtained by the inverse STFT of $\hat{\mathbf{c}}_j(n,f)$.

B. NMF-based source variance model

NMF has been a well-known technique for latent matrix factorization [19] and shown to be powerful in modeling audio spectra [6], [20]. It has been widely applied to single channel audio source separation where the mixture spectrogram is usually factorized into two latent matrices characterizing the spectral basis and the time activation [20]. When adapting NMF to the considered LGM summarized in Section II-A, the nonnegative source variances $v_j(n,f)$ can be approximated as

$$v_j(n,f) = \sum_{k=1}^{K_j} w_{jfk}h_{jkn}, \quad (10)$$

where w_{jfk} is an entry of the spectral basis matrix $\mathbf{W}_j \in \mathbb{R}_+^{F\times K_j}$, h_{jkn} is an entry of the activation matrix $\mathbf{H}_j \in \mathbb{R}_+^{K_j\times N}$, and K_j the number of latent components in the NMF model.

To our best knowledge, this NMF formulation for the source variances within the LGM was first presented in [38], and then further discussed in [36], [37]. However, in those works, the basis matrix $\mathbf{W}_j$ is not a GSSM as proposed in this article (presented in Section III-A), and thus the parameters $\{\mathbf{W}_j, \mathbf{H}_j\}$ were estimated differently.

C. Estimation of the model parameters

The set of parameters θ is estimated by minimizing the criterion (8) using a generalized EM algorithm (GEM) [45]. This algorithm consists in alternating between E step and M step. In the E step, given the observed empirical covariance matrix $\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)$ and the current estimate of θ, the conditional expectation of the natural statistics is computed as [31]

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_j(n,f) &= \mathbf{G}_j(n,f)\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)\mathbf{G}_j^H(n,f) + \\ &\quad (\mathbf{I} - \mathbf{G}_j(n,f))\mathbf{\Sigma}_j(n,f), \end{aligned} \quad (11)$$

where $\mathbf{G}_j(n,f) = \mathbf{\Sigma}_j(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)$ is the Wiener gain, $\mathbf{I}$ is an $I\times I$ identity matrix. Then in the M step, given $\widehat{\mathbf{\Sigma}}_j(n,f)$ the parameters $\theta_j = \{v_j(n,f), \mathbf{R}_j(f)\}_{n,f}$ associated to each j-th source are updated in the maximum likelihood sense by optimizing the following criterion [34]:

$$\mathcal{L}(\theta_j) = \sum_{n,f} \mathrm{tr}\big(\mathbf{\Sigma}_j^{-1}(n,f)\widehat{\mathbf{\Sigma}}_j(n,f)\big) + \log\det\big(\pi\mathbf{\Sigma}_j(n,f)\big). \quad (12)$$

By computing the derivatives of $\mathcal{L}(\theta_j)$ with respect to $v_j(n,f)$ and each entry of $\mathbf{R}_j(f)$ and equating them to zero, the iterative updates for these parameters are found as

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f) \quad (13)$$

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f)) \quad (14)$$

At each EM iteration, once $v_j(n,f)$ is updated in the M step by (14), it will be further constrained by NMF as (10). For this purpose, given the matrix of the current source variance estimate $\mathbf{V}_j \in \mathbb{R}_+^{F\times N}$ whose entries are $v_j(n,f)$, the corresponding NMF parameters are estimated by minimizing the Itakura-Saito divergence, which offers scale-invariant property, as

$$\underset{\mathbf{H}_j\geq 0, \mathbf{W}_j\geq 0}{min}\; D(\mathbf{V}_j\|\mathbf{W}_j\mathbf{H}_j), \quad (15)$$

where $D(\mathbf{V}_j\|\mathbf{W}_j\mathbf{H}_j) = \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}\big(v_j(n,f)\|w_{jfk}h_{jkn}\big)$, and

$$d_{IS}(x\|y) = \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1. \quad (16)$$

The parameters $\{\mathbf{W}_j, \mathbf{H}_j\}$ are usually initialized with random non-negative values and are iteratively updated via the well-known multiplicative update (MU) rules [19], [20].

III. PROPOSED GSSM-BASED MULTICHANNEL APPROACH

The global workflow of the proposed approach is depicted in Fig. 1. In the following, we will first review a training phase for the GSSM construction based on NMF in Section III-A. We then propose the NMF-based source variance model fitting with sparsity constraint in Section III-B. Finally, we derive the generalized EM algorithm for the parameter estimation in Section III-C. Note that we focus on NMF as the spectral model in this paper, however, the whole idea of the proposed approach can potentially be used for other spectral models than NMF such as GMM or DNN.

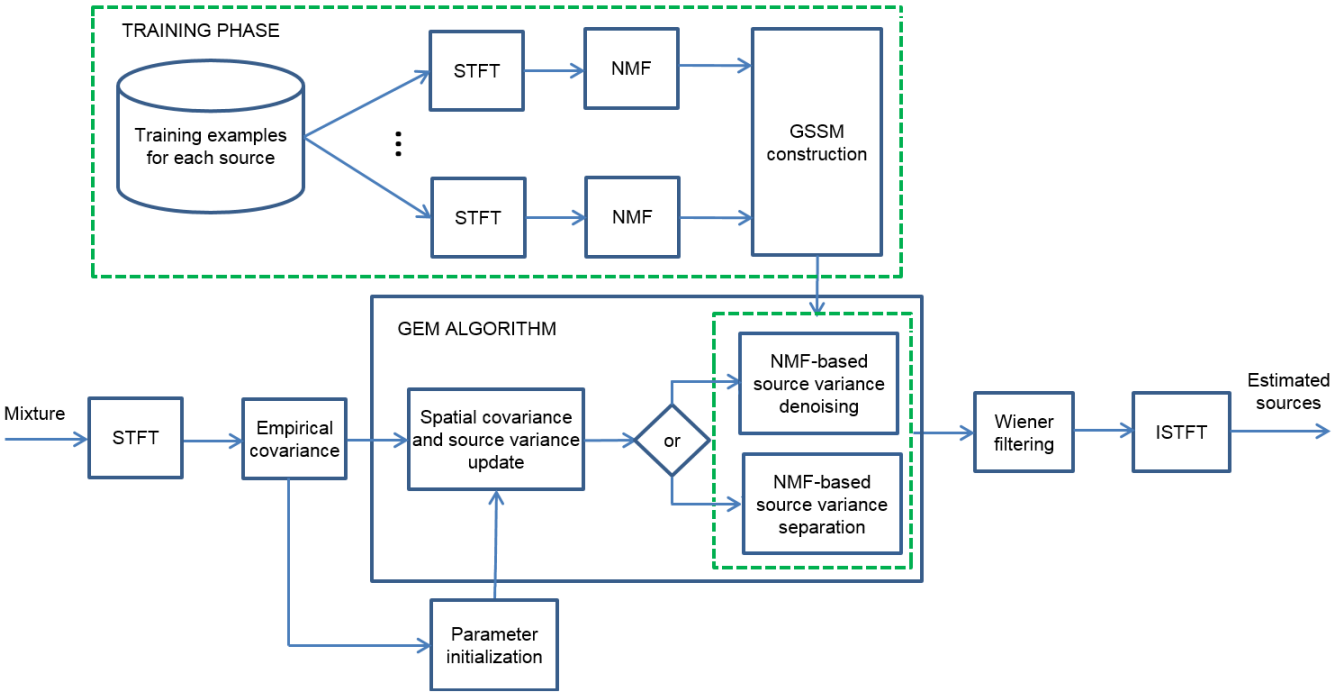

Fig. 1. General workflow of the proposed source separation approach. The top green dashed box describes the training phase for the GSSM construction. Bottom blue boxes indicate processing steps for source separation. Green dashed boxes indicate the novelty compared to the existing works [36]–[38].

A. GSSM construction

In this section, we review the GSSM construction, which was introduced in [14], [17]. We assume that the types of sources in the mixture are known and some recorded examples of such sounds are available. This is actually feasible in practice. For instance, in the speech enhancement, one target source is speech and another is noise and one can easily find speech and noise recordings. We need several examples for each type of source as one recording is usually not fully representative of the others and a source like "noise" is poorly defined. Let us denote by $s_j^l(t)$ a l-th single-channel learning example of j-th source and its corresponding spectrogram obtained by STFT $\mathbf{S}_j^l$. First, $\mathbf{S}_j^l$ is used to learn the corresponding NMF spectral dictionary, denoted by $\mathbf{W}_j^l$, by optimizing the similar criterion as (15):

$$\underset{\mathbf{H}_j^l \geq 0, \mathbf{W}_j^l \geq 0}{min} D(\mathbf{S}_j^l \| \mathbf{W}_j^l \mathbf{H}_j^l) \tag{17}$$

where $\mathbf{H}_j^l$ is the time activation matrix. Given $\mathbf{W}_j^l$ for all examples $l = 1, ..., L_j$ of the j-th source, the GSSM for the j-t source is constructed as

$$\mathbf{U}_j = [\mathbf{W}_j^1, \ldots, \mathbf{W}_j^{L_j}], \tag{18}$$

then the GSSM for all the sources is computed by

$$\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_J]. \tag{19}$$

As an example for speech and noise separation, in the practical implementation, we may need several speech examples for different male voices and female voices (*e.g.,* 5 examples in total), and examples of different types of noise such as those from outdoor environment, cafeteria, waterfall, street, etc.,. (*e.g.,* 6 examples in total).

Note that as another variant investigated in this work, the GSSM $\mathbf{U}_j$ can be constructed differently by first concatenating all examples for each source ($\mathbf{S}_j = [\mathbf{S}_j^1, \ldots, \mathbf{S}_j^{L_j}]$), and then performing NMF on the concatenated spectrogram only once by optimizing the criterion

$$\underset{\mathbf{H}_j \geq 0, \mathbf{U}_j \geq 0}{min} D(\mathbf{S}_j \| \mathbf{U}_j \mathbf{H}_j). \tag{20}$$

We will show in the experiment that this way of constructing the GSSM does not provide as good source separation performance as the one presented before by (18).

B. Proposed source variance fitting with GSSM and mixed group sparsity constraint

As the GSSM is constructed to guide the NMF-based source variance constraint, we propose two fitting strategies as follows:

1) Source variance denoising: Motivated by the source variance model (10), when exploiting the GSSM model we propose a variant as

$$v_j(n,f) = \sum_{k=1}^{P_j} u_{jfk} \tilde{h}_{jkn}, \tag{21}$$

where u_{jfk} is an entry of $\mathbf{U}_j$, $\tilde{h}_{jkn}$ is an entry of the corresponding activation matrix $\widetilde{\mathbf{H}}_j \in \mathbb{R}_+^{P_j \times N}$. This leads to a straightforward extension of the conventional optimization criterion described by (15) where $\widetilde{\mathbf{H}}_j$ is now estimated by optimizing the criterion:

$$\underset{\widetilde{\mathbf{H}}_j \geq 0}{min} D(\mathbf{V}_j \| \mathbf{U}_j \widetilde{\mathbf{H}}_j) + \lambda \Omega(\widetilde{\mathbf{H}}_j), \tag{22}$$

where $\mathbf{U}_j$ is constructed by (18) or (20) and fixed, $\Omega(\widetilde{\mathbf{H}}_j)$ presents a penalty function imposing sparsity on $\widetilde{\mathbf{H}}_j$, and λ is a trade-off parameter determining the contribution of the penalty. Note that as the GSSM $\mathbf{U}_j$ constructed in (18) becomes a large matrix when the number of examples L_j for each source increases, and it is actually a redundant dictionary since different examples may share similar spectral patterns. Thus to fit the source variances with the GSSM, sparsity constraint is naturally needed in order to activate only a subset of $\mathbf{U}_j$ which represents the spectral characteristics of the sources in the mixture [46]–[48].

2) Source variance separation: We propose another source variance model as

$$v(n,f) = \sum_{k=1}^{K} u_{fk}\tilde{h}_{kn}, \tag{23}$$

where $v(n,f) = \sum_{j=1}^{J} v_j(n,f)$, u_{fk} is an entry of the GSSM model $\mathbf{U}$ constructed as (19) and fixed, $K = \sum_{j=1}^{J} P_j$. Under this model, let $\widetilde{\mathbf{V}} = \sum_{j=1}^{J} \mathbf{V}_j$ be the matrix of the total source variance estimate, it is then decomposed by solving the following optimization problem

$$\underset{\widetilde{\mathbf{H}}\geq 0}{min}\ D(\widetilde{\mathbf{V}}\|\mathbf{U}\widetilde{\mathbf{H}}) + \lambda\Omega(\widetilde{\mathbf{H}}) \tag{24}$$

where $\Omega(\widetilde{\mathbf{H}})$ presents a penalty function imposing sparsity on the activation matrix $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{H}}_1^\top, ..., \widetilde{\mathbf{H}}_J^\top]^\top \in \mathbb{R}_+^{K\times N}$ the total number of rows in $\widetilde{\mathbf{H}}$. This criterion can be seen as an additional NMF-based separation step applied on the source variances, while criterion (22) and other existing works [36]–[38] do not perform any additional separation of the variances, but more like denoising of the already separated variances. For the sake of simplicity, in the following, we only present the algorithm derivation for the criterion (24), but a strong synergy can be found for the criterion (22).

Recent works in audio source separation have considered two penalty functions, namely *block* sparsity-inducing penalty [14] and *component* sparsity-inducing penalty [17]. The former one enforces the activation of *relevant examples* only while omitting irrelevant ones since their corresponding activation block in $\widetilde{\mathbf{H}}$ will likely converge to zero. The latter one, on the other hand, enforces the activation of *relevant components* in $\mathbf{U}$ only. It is motivated by the fact that only a part of the spectral model learned from an example may fit well with the targeted source in the mixture, while the remaining components in the model do not. Thus instead of activating the whole block, the component sparsity-inducing penalty allows selecting only the more likely relevant spectral components from $\mathbf{U}$. Inspired by the advantage of these penalty functions, in our recent work we proposed to combine them in a more general form as [18]

$$\Omega(\widetilde{\mathbf{H}}) = \gamma\sum_{p=1}^{P}\log(\epsilon + \|\mathbf{H}_p\|_1) + (1-\gamma)\sum_{k=1}^{K}\log(\epsilon + \|\mathbf{h}_k\|_1), \tag{25}$$

where the first term on the right hand side of the equation presents the block sparsity-inducing penalty, the second term presents the component sparsity-inducing penalty, and $\gamma \in [0,1]$ weights the contribution of each term. In (25), $\mathbf{h}_k \in \mathbb{R}_+^{1\times N}$ is a row (or component) of $\widetilde{\mathbf{H}}$, $\mathbf{H}_p$ is a subset of $\widetilde{\mathbf{H}}$ representing the activation coefficients for p-th block, P is the total number of blocks, ϵ is a non-zero constant, and $\|.\|_1$ denotes ℓ_1-norm operator. In the considered setting, a block represents one training example for a source and P is the total number of used examples (*i.e.,* $P = \sum_{j=1}^{J} L_j$).

By putting (25) into (24), we now have a complete criterion for estimating the activation matrix $\widetilde{\mathbf{H}}$ given $\widetilde{\mathbf{V}}$ and the pre-trained spectral model $\mathbf{U}$. The derivation of MU rule for updating $\widetilde{\mathbf{H}}$ is presented in the Appendix.

C. Proposed multichannel algorithm

Within the LGM, a generalized EM algorithm used to estimate the parameters $\{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ by considering the set of hidden STFT coeffients of all the source images $\{\mathbf{c}_j(n,f)\}_{n,f}$ as the *complete data*. The overview for the GEM derivation are presented in Section II-C, and more details can be found in [34], [37].

For the proposed approach as far as the GSSM concerned, the E-step of the algorithm remains the same as in [34]. In the M-step, we additionally perform the optimization defined either by (22) (for source variance denoising) or by (24) (for source variance separation). This is done by the MU rules so that the estimated intermediate source variances $v_j(n,f)$ are further updated with the supervision of the GSSM. The detail of overall proposed algorithm with source variance separation is summarized in Algorithm 1.

Note that this generalized EM algorithm requires the same order of computation compared to the existing method [37], [38] as sparsity constraint and bigger GSSM size does not significantly affect the overall computational time. As an example, for separating a 10-second long mixture presented in our experiment, both [38] and our proposed method (when non-optimally implemented in Matlab) take about 400 seconds when running in a laptop with Intel Core i5 Processor, 2.2 GHz, and 8 GB RAM.

IV. EXPERIMENTS

A. Dataset and parameter settings

We validated the performance of the proposed approach in an important speech enhancement use case where we know already two types of sources in the mixture: speech and noise. For a better comparison with the state of the art, we used the benchmark development dataset of the "Two-channel mixtures of speech and real-world background noise" (BGN) task[1] within the SiSEC 2016 [4]. This devset contains stereo mixtures of 10-second duration and 16 kHz sampling rate. They were the mixture of male/female speeches and real-world noises recorded from different public environments: cafeteria (Ca), square (Sq), and subway (Su). Overall there were nine mixtures: three with Ca noise, four with Sq noise, and two with Su noise. The signal-to-noise ratio was drawn randomly between -17 and +12 dB by the dataset creators.

[1]https://sisec.inria.fr/sisec-2016/bgn-2016/

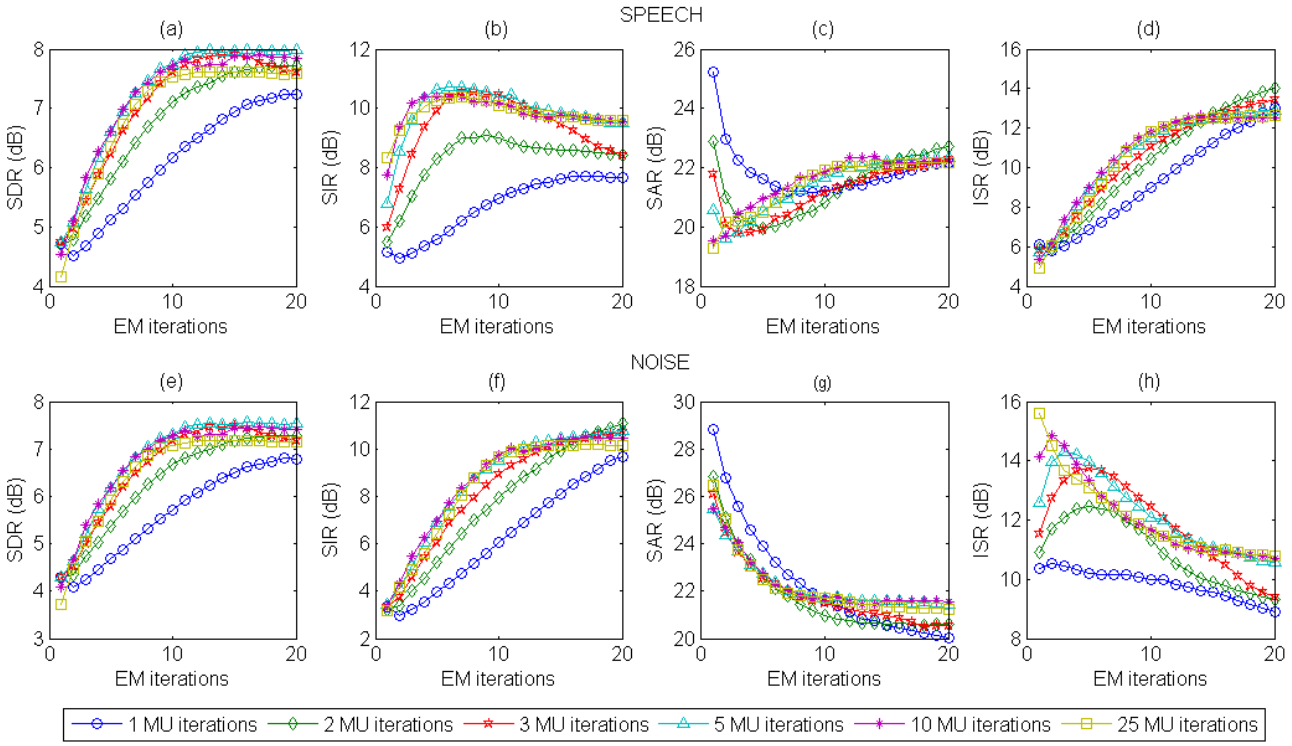

Fig. 2. Average separation performance obtained by the proposed method over stereo mixtures of speech and noise as functions of EM and MU iterations. (a): speech SDR, (b): speech SIR, (c): speech SAR, (d): speech ISR, (e): noise SDR, (f): noise SIR, (g): noise SAR, (h): noise ISR

Our works in single-channel case [16], [18] and preliminary tests on multichannel case show that only a few examples for each source could be enough to train an efficient GSSM. Thus, for training the generic speech spectral model, we took only one male voice and two female voices from the SiSEC 2015[2]. These three speech examples are also 10-second length. We performed the listening check to confirm that these examples used for the speech and noise model training are different from those in the devset, which were used for testing. For training the generic noise spectral model, we extracted five noise examples from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND)[3]. Again they were 10-second length and contained three types of environmental noise: cafeteria, square, metro. The STFT window length was 1024 for all train and test files. The number of NMF components in $\mathbf{W}_j^l$ for each speech example was set to 32, while that for noise example was 16. These values were found to be reasonable in [15] and our work on single-channel case [18]. Each $\mathbf{W}_j^l$ were obtained by optimizing (17) with 20 MU iterations.

Initialization of the spatial covariance matrices: As suggested in [34], we firstly tried to initialize the spatial covariance matrix $\mathbf{R}_j(f)$ by performing hierarchical clustering on the mixture STFT coefficients $\mathbf{x}(n,f)$. But this strategy did not give us a good separation performance as the noise source in the considered mixtures is diffuse (*i.e.,* it does not come from a single direction). Thus we initialized the noise spatial covariance matrix based on the diffuse model where noise is assumed to come uniformly from all spatial directions. With this assumption, the diagonal entries of the noise spatial covariance matrix are one and the off-diagonal entries are real-valued computed as in [49]

$$r_{1,2}(f) = r_{2,1}(f) = \frac{\sin(2\pi f d/v)}{2\pi f d/v}, \tag{26}$$

where d is the distance between two microphones and $v = 334$ m/s the sound velocity. The spatial covariance matrix for the speech source was initialized by the full-rank direct+diffuse model detailed in [34] where the speech's direction of arrival (DoA) was set to 90 degrees. This DoA initialization was chosen for balancing the fact that the speech direction can vary between 0 degree and 180 degrees in each mixture and we did not have access to the ground truth information while performing the test.

The source separation performance for all approaches was evaluated by two sets of criteria. The four power-based criteria: the signal to distortion ratio (SDR), the signal to interference ratio (SIR), the signal to artifacts ratio (SAR), and the source image to spatial distortion ratio (ISR), measured in dB where the higher the better [50]. The four perceptually-motivated criteria: the overall perceptual score (OPS), the target-related perceptual score (TPS), the artifact-related perceptual score (APS), and the interference-related perceptual score (IPS) [51], where a higher score is better. As power-based criteria are more widely used in source separation community, the hyper-parameters for each algorithm were chosen in order to

[2]https://sisec.inria.fr/sisec-2015/2015-underdetermined-speech-and-music-mixtures/.

[3]http://parole.loria.fr/DEMAND/.

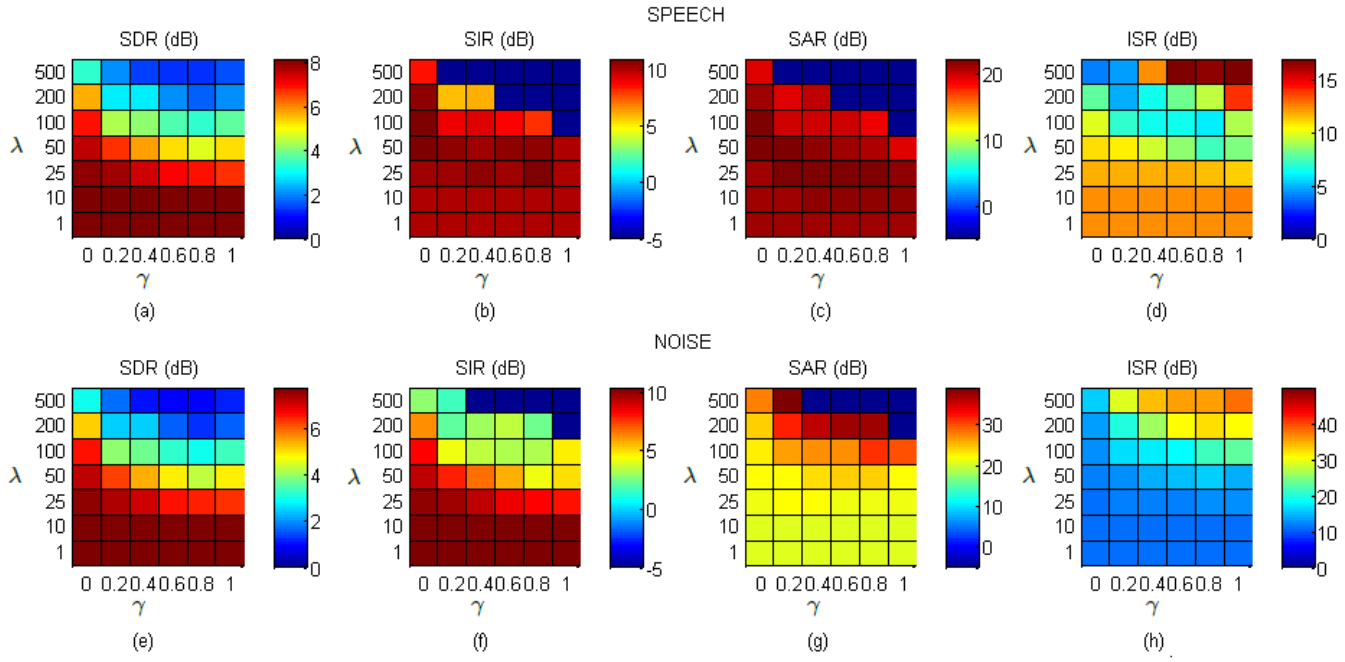

Fig. 3. Average separation performance obtained by the proposed method over stereo mixtures of speech and noise as functions of λ and γ. (a): speech SDR, (b): speech SIR, (c): speech SAR, (d): speech ISR, (e): noise SDR, (f): noise SIR, (g): noise SAR, (h): noise ISR

maximize the SDR - the most important metric as it reflects the overall signal distortion.

B. Algorithm analysis

1) Algorithm convergence: separation results as functions of EM and MU iterations: We first investigate the convergence in term of separation performance of the derived Algorithm 1 by varying the number of EM and MU iterations and computing the separation results obtained on the benchmark BGN dataset. In this experiment, we set $\lambda = 10$ and $\gamma = 0.2$ as we will show in next section that these values offer both the stability and the good separation performance. The speech and noise separation results, measured by the SDR, SIR, SAR, and ISR, averaged over all mixtures in the dataset, illustrated as functions of the EM and MU iterations, are shown in Fig. 2.

As it can be seen, generally the SDR increases when the number of EM and MU iterations increases. With 10 or 25 MU iterations, the algorithm converges nicely and saturates after about 10 EM iterations. The best separation performance was obtained with 10 MU iterations and 15 EM iterations. It is also interesting to see that with a small number of MU iterations like 1, 2, or 3, the separation results are quite poor and the algorithm is less stable as it varies significantly even with a large number of EM iterations. This reveals the effectiveness of the proposed NMF constraint (24).

2) Separation results with different choices of λ and γ: We further investigate the sensitivity of the proposed algorithm to two parameters λ and γ, which determine the contribution of sparsity penalty to the NMF constraint in (24). For this purpose, we varied the values of these parameters, $\lambda = \{1, 10, 25, 50, 100, 200, 500\}$, $\gamma = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, and applied the corresponding source separation algorithm presented in the Algorithm 1 on the benchmark BGN dataset. The number of EM and MU iterations are set to 15 and 10, respectively, as these values guarantee the algorithm's convergence shown in Fig. 2. The speech and noise separation results, measured by the SDR, SIR, SAR, and ISR, averaged over all mixtures in the dataset, represented as functions of λ and γ, are shown in Fig. 3.

It can be seen that the proposed algorithm is less sensitive to the choice of γ, while more sensitive to the choice of λ, and the separation performance greatly decreases with $\lambda > 10$. The best choice for these parameters in term of the SDR are $\lambda = 10, \gamma = 0.2$. With the small value of λ (*e.g.,* $\lambda = 1$), varying γ does not really affect the separation performance as the evaluation criteria are quite stable. We noted that with $\gamma = 0.2$, the algorithm offers 0.2 dB and 1.0 dB SDR, which are higher than when $\gamma = 0$ and $\gamma = 1$, respectively. This confirms the effectiveness of the mixed sparsity penalty (25) in the multichannel setting.

C. Comparison with the state of the art

We compare the speech separation performance obtained on the BGN dataset of the proposed approach with its close prior art (*i.e.* Arberet's algorithm [38]) and other state-of-the-art methods presented at the SiSEC campaign over different years since 2013. The results of these methods were submitted by the authors and evaluated by the SiSEC organizers [4], [52], [53]. All comparing methods are summarized as follows:

- Martinez-Munoz's method (SiSEC 2013) [52]: this algorithm exploits source-filter model for the speech source and the noise source is modeled as a combination of pseudo-stationary broadband noise, impulsive noise, and

Algorithm 1 Proposed GSSM + SV separation algorithm

Require:

Mixture signal $\mathbf{x}(t)$

List of examples of each source in the mixture $\{s_j^l(t)\}_{j=1:J,l=1:L_j}$

Hyper-parameters λ, γ, MU-iteration

Ensure: Source images $\hat{\mathbf{c}}_j(t)$ separated from $\mathbf{x}(t)$

- Compute the mixture STFT coefficients $\mathbf{x}(n,f) \in \mathbb{C}^{F\times N}$ and then $\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f) \in \mathbb{C}^{I\times I}$ by (7)
- Construct the GSSM model $\mathbf{U}_j$ by (18), then $\mathbf{U} \in \mathbb{R}_+^{F\times K}$ by (19)
- Initialize the spatial covariance matrices $\mathbf{R}_j(f), \forall j, f$ (see Section IV-A)
- Initialize the non-negative time activation matrix for each source $\widetilde{\mathbf{H}}_j$ randomly, then $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{H}}_1^\top, \ldots, \widetilde{\mathbf{H}}_J^\top]^v \in \mathbb{R}_+^{K\times N}$
- Initialize the source variance $v_j(n,f) = [\mathbf{U}_j \widetilde{\mathbf{H}}_j]_{n,f}$

// Generalized EM algorithm for the parameter estimation:

repeat

// E step (perform calculation for all j, n, f):

$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)$ // eq. (4)

$\mathbf{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^J v_j(n,f)\,\mathbf{R}_j(f)$ // eq. (5)

$\mathbf{G}_j(n,f) = \mathbf{\Sigma}_j(n,f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)$ // Wiener gain

$\widehat{\mathbf{\Sigma}}_j(n,f) = \mathbf{G}_j(n,f)\widehat{\mathbf{\Psi}}_{\mathbf{x}}(n,f)\mathbf{G}_j^H(n,f) + (\mathbf{I} - \mathbf{G}_j(n,f))\mathbf{\Sigma}_j(n,f)$ // eq. (11)

// M step: updating spatial covariance matrix and unconstrained source spectra

$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^N \frac{1}{v_j(n,f)}\widehat{\mathbf{\Sigma}}_j(n,f)$ // eq. (13)

$v_j(n,f) = \frac{1}{I}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{\Sigma}}_j(n,f))$ // eq. (14)

$\mathbf{V}_j = \{v_j(n,f)\}_{n,f}$

$\widetilde{\mathbf{V}} = \sum_{j=1}^J \mathbf{V}_j$

// MU rules for NMF inside M step to further constrain source spectra by the GSSM

for $iter = 1, ..., $MU-iteration **do**

for $p = 1, ..., P$ **do**

$\mathbf{Y}_p \leftarrow \frac{1}{\epsilon + \|\mathbf{H}_p\|_1}$

end for

$\mathbf{Y} = [\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_P^\top]^\top$

for $k = 1, ..., K$ **do**

$\mathbf{z}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_k\|_1}$

end for

$\mathbf{Z} = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_K^\top]^\top$

// Updating activation matrix

$\widehat{\mathbf{V}} = \mathbf{U}\widetilde{\mathbf{H}}$

$\widetilde{\mathbf{H}} \leftarrow \widetilde{\mathbf{H}} \odot \left(\frac{\mathbf{U}^\top(\widetilde{\mathbf{V}} \odot \widehat{\mathbf{V}}^{\cdot -2})}{\mathbf{U}^\top(\widehat{\mathbf{V}}^{\cdot -1}) + \lambda(\gamma\mathbf{Y} + (1-\gamma)\mathbf{Z})}\right)^{\cdot \frac{1}{2}}$ // eq. (31)

end for

$v_j(n,f) = [\mathbf{U}_j\widetilde{\mathbf{H}}_j]_{n,f}$ // updating constrained spectra

until convergence

- Source separation by multichannel Wiener filtering (9)
- Time domain source images $\hat{\mathbf{c}}_j(t)$ are obtained by the inverse STFT of $\hat{\mathbf{c}}_j(n,f)$.

pitched interferences. The parameter estimation is based on the MU rules employed in non-negative matrix factorization.

- Wang's method [54] (SiSEC 2013): this algorithm performs well-known frequency domain independent component analysis (ICA). The associated permutation problem is solved by a novel region-growing permutation alignment technique.
- Le Magoarou's method [9] (SiSEC 2013): this approach uses text transcript of the speech source in the mixture as prior information to guide the source separation process. The algorithm is based on the nonnegative matrix partial co-factorization.
- Bryan's method [55] (SiSEC 2013): this interactive approach exploits human annotation on the mixture spectrogram to guide and refine the source separation process. The modeling is based on the probabilistic latent component analysis (PLCA), which is equivalent to NMF.
- Rafii's method [56] (SiSEC 2013): this technique uses a similarity matrix to separate the repeating background from the non-repeating foreground in a mixture. The underlying assumption is that the background is dense and low-ranked, while the foreground is sparse and varied.
- Ito's method [57] (SiSEC 2015): this is a permutation-free frequency-domain blind source separation algorithm via full-band clustering of the time-frequency (T-F) components. The clustering is performed via MAP estimation of the parameters with EM algorithm.
- Liu's method [4] (SiSEC 2016): the algorithm performs Time Difference of Arrival (TDOA) clustering based on GCC-PHAT.
- Wood's method [58] (SiSEC 2016): this recently proposed algorithm first applies NMF to the magnitude spectrograms of the mixtures with channels concatenated in time. Each dictionary atom is clustered to either the speech or the noise according to its spatial origin.
- Arberet's method [38]: using the similar local Gaussian model, the algorithm further constrains the intermediate source variances by unsupervised NMF with criterion (15). Such algorithm is implemented by Ozerov *et. al.* in [37]. This method is actually the most relevant prior art to compare with as it falls in the same LGM framework.

The proposed approach with different variants are summarized as:

- GSSM + SV denoising: The proposed GSSM + full-rank spatial covariance approach where the estimated variances of each sources $\mathbf{V}_j$ are further constrained by criterion (22). We submitted results obtained by this method to the SiSEC 2016 BGN task and obtained the best performance over the actual test set in term of SDR [4].
- GSSM + SV separation: The proposed approach with source variance separation by optimizing criterion (24). In order to investigate the benefit of the sparsity constraint, we further report the results obtained by this method when $\lambda = 0$. Finally, to confirm the effectiveness of the GSSM construction by (18), we report the results obtained when the GSSM of the same size is learned

jointly by concatenating all example's spectrograms $\mathbf{S}_j^l$ as (20). In this case, only the component sparsity is applied (*i.e.,* $\gamma = 0$) as block does not exist. This setting is named "GSSM'+component sparsity" in Table 1.

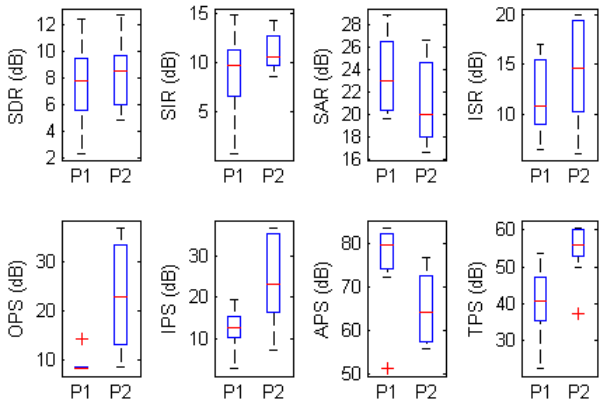

Fig. 4. Boxplot for the speech separation performance obtained by the proposed "GSSM + SV denoising" (P1) and "GSSM + SV separation" (P2) methods.

The separation results obtained by different methods for each noisy environment (Ca, Sq, Su), and the average overall mixtures are summarized in Table 1. The boxplot to illustrate the variance of the results obtained by the two proposed approaches is shown in Fig. 4. It is interesting to see that the results obtained by the proposed approach without sparsity constraint were lower than that of Arberet's method for all noisy environments, even the former used the pre-trained GSSM while the latter is completely unsupervised. It reveals that the GSSM itself is redundant and contains some irrelevant spectral patterns with the actual sources in the mixture. Thus constraining the source variances by the GSSM without a relevant spectral pattern selection guided by the sparsity penalty is even worse than the unsupervised NMF case where the spectral patterns are randomly initialized and then updated by MU rules. The importance of such sparsity penalty is explicitly confirmed by the fact that the results obtained by the proposed approach with sparsity constraint are far better than the setting without the sparsity constraint. Also, it is not surprising to see that the "GSSM + SV denoising" clearly outperforms Arberet's method (except for the ISR and the TPS) in all noisy environments as the former exploits additional information about the types of sources in the mixtures in order to learn the GSSM in advance. The "GSSM + SV separation" offers better separation performance in terms of SDR, SIR, OPS, IPS, on square and subway environments as well as on average compared to the "GSSM + SV denoising" and the "GSSM' + component sparsity". This confirms the effectiveness of the proposed source variance separation criterion (24) and the GSSM construction (18).

When compared to the top-performing state-of-the-art methods in the SiSEC campaigns, the proposed approach performs generally better in terms of the energy-based criteria but worse for the perceptually-motivated ones. Especially in Ca environment the OPS obtained by the proposed approach is far below those offered by other methods. This may be due to the fact that the hyper-parameters were optimized for the SDR, but not the OPS. The "GSSM + SV separation" with sparsity constraint outperforms all other methods, but Wang's approach, in terms of the SDR, the most important energy-based criterion, at all noisy environment. This confirms the effectiveness of the proposed approach where the GSSM is successfully exploited in the LGM framework. It should be noted that Wang's method [54] is based on the frequency-domain ICA so it is not applicable for under-determined mixtures where the number of sources is larger than the number of channels. Also, in this method, an additional post-filtering technique was applied to the separated speech source so as to maximize the denoising capability.

V. CONCLUSION

In this paper, we have presented a novel multichannel audio source separation algorithm weakly guided by some source examples. The considered approach exploits the use of generic source spectral model learned by NMF within the well-established local Gaussian model. In particular, we have proposed a new source variance separation criterion in order to better constrain the intermediate source variances estimated in each EM iteration. Experiments with the benchmark dataset from the SiSEC campaigns have confirmed the effectiveness of the proposed approach compared to the state of the art. Motivated by the effectiveness of the GSSM, future work can be devoted to extending the current approach in order to exploit in addition the use of a *generic spatial covariance model*, which remains to be defined. In addition, the theoretical grounding of the source variance separation criterion needs to be further investigated. Another promising investigation could be extending the idea of source variance separation to DNN-based models inspired by the work of Nugraha *et al.* [24].

VI. ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers of this manuscript and [43] for their helpful and constructive comments that greatly contributed to improving the quality of the paper. We also would like to thank Professor Nobutaka Ono for providing us results of the SiSEC 2013 and the SiSEC 2015.

APPENDIX
DERIVATION OF MU RULE FOR UPDATING THE ACTIVATION MATRIX IN ALGORITHM1

Let $\mathcal{L}(\widetilde{\mathbf{H}})$ denote the minimization criterion (24) with the mixed sparsity constrained $\Omega(\widetilde{\mathbf{H}})$ defined as in (25) and $D(\cdot\|\cdot)$ being IS divergence. The partial derivative of $\mathcal{L}(\widetilde{H})$ with respect to an entry h_{kn} is

$$\nabla_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) = \sum_{f=1}^{F} u_{fk}\left(\frac{1}{[\mathbf{U}\widetilde{\mathbf{H}}]_{n,f}} - \frac{v(n,f)}{[\mathbf{U}\widetilde{\mathbf{H}}]_{n,f}^2}\right) + \frac{\lambda.\gamma}{\epsilon + \|\mathbf{H}_p\|_1} + \frac{\lambda.(1-\gamma)}{\epsilon + \|\mathbf{h}_k\|_1} \quad (27)$$

Methods	Ca1				Sq1				Su1				Average			
	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR	SDR	SIR	SAR	ISR
	OPS	IPS	APS	TPS	OPS	IPS	APS	TPS	OPS	IPS	APS	TPS	OPS	IPS	APS	TPS
Martinez-Munoz*	5.4	15.4	6.1	-	9.6	17.3	10.7	-	1.5	5.8	5.8	-	6.4	14.1	7.9	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Wang* [54]	10.4	21.6	12.8	13.5	10.3	19.1	12.3	15.0	8.1	19.3	10.0	10.7	**9.8**	**20.0**	**12.0**	**13.5**
	41.9	60.4	52.9	58.8	37.4	51.8	53.3	55.7	33.1	52.4	42.3	43.7	**37.9**	**54.8**	**50.7**	**54.1**
Le Magoarou* [9]	9.2	11.6	13.4	19.8	4.0	6.2	8.3	20.4	-5.2	-4.5	2.7	9.7	3.7	5.6	8.8	17.8
	31.3	29.3	52.8	77.9	38.9	45.2	39.9	75.6	22.9	24.6	34.8	28.3	32.8	35.3	43.1	65.9
Bryan* [55]	5.6	18.4	5.9	-	10.2	15.6	12.1	-	4.2	13.6	4.9	-	7.3	16.1	7.6	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Rafii* [56]	8.8	13.0	12.1	13.3	6.2	9.6	8.9	10.7	-2.7	-2.7	4.4	11.0	5.1	8.0	9.0	11.6
	29.2	27.3	58.0	56.2	34.6	38.7	55.8	60.7	23.9	21.6	51.3	50.4	30.4	31.1	55.5	56.9
Ito* [57]	7.2	25.9	7.2	-	8.9	23.7	9.1	-	4.9	15.3	5.6	-	7.4	22.6	7.7	-
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Liu*	-1.0	4.9	19.7	4.1	-8.5	-2.9	15.1	1.9	-12.8	-8.0	7.6	3.8	-7.0	-1.4	15.0	3.1
	9.5	16.8	77.1	42.3	14.2	18.9	70.2	38.8	21.2	15.7	60.0	49.5	14.2	17.5	70.3	42.3
Wood* [58]	3.0	9.4	5.0	3.7	1.9	2.4	4.0	7.5	0.2	-2.6	1.3	2.5	1.9	3.6	3.7	5.1
	33.7	60.7	39.0	40.5	38.6	60.5	43.3	57.6	25.9	47.6	31.7	24.4	34.1	57.7	39.3	44.5
Arberet [37], [38]	9.1	10.0	16.1	19.5	3.3	3.3	10.4	15.3	-0.2	-1.2	9.5	11.7	4.4	4.6	12.1	**15.9**
	13.3	10.9	70.4	50.5	8.3	10.5	82.3	47.5	10.2	3.7	56.6	23.4	10.4	9.1	72.6	43.2
GSSM + SV denoising	10.5	11.8	27.7	16.2	7.0	8.5	22.0	9.8	5.1	5.6	20.7	8.1	7.7	9.0	23.6	11.6
$(\lambda = 10, \gamma = 0.2)$	8.4	12.7	83.0	49.9	8.5	14.7	77.6	39.0	11.3	7.8	61.8	27.6	18.1	12.5	75.9	40.1
GSSM + SV separation	7.9	10.2	20.2	11.2	-1.1	-2.6	17.6	8.0	-1.6	-3.2	20.4	7.6	1.8	1.5	19.1	8.9
(No sparsity constraint)	25.0	19.3	64.1	55.8	32.4	29.4	55.0	60.2	18.7	11.4	56.0	35.7	18.8	22.0	58.2	53.3
GSSM + SV separation	7.3	10.0	19.4	9.7	4.4	6.1	16.0	6.9	2.4	1.8	18.3	8.8	4.9	6.5	17.7	8.3
(GSSM' + component sparsity)	20.7	18.2	66.2	48.1	30.3	28.3	57.9	52.8	21.6	16.5	56.0	43.0	21.3	22.3	60.3	49.1
GSSM + SV separation	10.6	13.5	25.6	19.6	7.8	11.1	19.3	12.3	5.0	7.1	18.7	9.5	**8.1**	**11.0**	**21.3**	**14.1**
$(\lambda = 10, \gamma = 0.2)$	11.4	13.0	81.6	61.0	31.6	31.4	62.0	57.4	23.7	27.8	47.3	37.6	**23.1**	**24.5**	**65.2**	**54.2**

TABLE I

SPEECH SEPARATION PERFORMANCE OBTAINED ON THE DEVSET OF THE BGN TASK OF THE SISEC CAMPAIGN. * INDICATES SUBMISSIONS BY THE AUTHORS AND "-" INDICATES MISSING INFORMATION.

This $\nabla_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}})$ can be written as a sum of two nonnegative parts, denoted by $\nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) \geq 0$ and $\nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) \geq 0$, respectively, as

$$\nabla_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) = \nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) - \nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) \tag{28}$$

with

$$\begin{aligned}\nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) &\triangleq \sum_{f=1}^{F} u_{fk}\frac{1}{[\mathbf{U}\widetilde{\mathbf{H}}]_{n,f}} + \frac{\lambda.\gamma}{\epsilon + \|\mathbf{H}_p\|_1} + \frac{\lambda.(1-\gamma)}{\epsilon + \|\mathbf{h}_k\|_1},\\ \nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}}) &\triangleq \sum_{f=1}^{F} u_{fk}\frac{v(n,f)}{[\mathbf{U}\widetilde{\mathbf{H}}]^2_{n,f}}.\end{aligned} \tag{29}$$

Following a standard approach for MU rule derivation [19], [20]), h_{kn} is updated as

$$h_{kn} \leftarrow h_{kn}\left(\frac{\nabla^-_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}})}{\nabla^+_{h_{kn}}\mathcal{L}(\widetilde{\mathbf{H}})}\right)^{.\eta}, \tag{30}$$

where $\eta = 0.5$ following the derivation in [47], [59], which was shown to produce an accelerated descent algorithm. Putting (29) into (30) and rewriting it in a matrix form, we obtain the updates of $\widetilde{\mathbf{H}}$ as

$$\widetilde{\mathbf{H}} \leftarrow \widetilde{\mathbf{H}} \odot \left(\frac{\mathbf{U}^\top(\widetilde{\mathbf{V}} \odot \widehat{\mathbf{V}}^{\cdot -2})}{\mathbf{U}^\top(\widehat{\mathbf{V}}^{\cdot -1}) + \lambda(\gamma\mathbf{Y} + (1-\gamma)\mathbf{Z})}\right)^{\cdot\frac{1}{2}}, \tag{31}$$

where $\widehat{\mathbf{V}} = \mathbf{U}\widetilde{\mathbf{H}}$, $\mathbf{Y} = [\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_P^\top]^\top$ with $\mathbf{Y}_p, p = 1, \ldots P$ an uniform matrix of the same size as $\mathbf{H}_p$ whose entries are $\frac{1}{\epsilon + \|\mathbf{H}_p\|_1}$, and $\mathbf{Z} = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_K^\top]^\top$ with $\mathbf{z}_k, k = 1, \ldots K$ a uniform vector of the same size as $\mathbf{h}_k$ whose entries are $\frac{1}{\epsilon + \|\mathbf{h}_k\|_1}$.

REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[3] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[4] 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 *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, 2017, pp. 323–332.

[5] A. Liutkus, J. L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Proc. IEEE Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.

[6] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[7] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, pp. 1775–1787, 2015.

[8] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.

[9] L. L. Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, pp. 1–5, 2014.

[10] N. Q. K. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," in *IEEE Int. Conf on Consumer Electronics (ICCE-Berlin)*, 2014, pp. 220–224.

[11] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 45–48.

[12] S. Ewert, B. Pardo, M. Mueller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[13] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "Motion informed audio source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[14] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[15] D. E. Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation - a novel user-friendly framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 261–272, 2017.

[16] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, and N. Q. K. Duong, "Single-channel speaker-dependent speech enhancement exploiting generic noise model learned by non-negative matrix factorization," in *Proc. IEEE Int. Conf. on Electronics, Information, and Communications (ICEIC)*, 2016, pp. 1–4.

[17] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.

[18] H. T. T. Duong, Q. C. Nguyen, C. P. Nguyen, T. H. Tran, and N. Q. K. Duong, "Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint," in *Proc. ACM Int. Sym. on Information and Communication Technology (SoICT)*, 2015, pp. 247–251.

[19] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural and Information Processing Systems 13*, 2001, pp. 556–562.

[20] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[21] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[22] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2135–2139.

[23] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[24] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 9, pp. 1652–1664, 2016.

[25] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[26] S. Michael, L. Jan, K. Ulrik, and C. Lucas, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Springer, 2007, pp. 1–34.

[27] S. R. A. Jourjine and O. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, June 2000, pp. 2985–2988.

[28] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[29] Z. El Chami, D. T. Pham, C. Serviere, and A. Guerin, "A new model based underdetermined source separation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008, pp. 147–150.

[30] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, article ID 24717, 2007.

[31] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[32] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[33] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1622–1637, 2016.

[34] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[35] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[36] M. Fakhry, P. Svaizer, and M. Omologo, "Audio source separation in reverberant environments using beta-divergence based nonnegative factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, 2017.

[37] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[38] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. IEEE Int. Conf. on Information Science, Signal Processing and their Applications (ISSPA)*, 2010, pp. 1–4.

[39] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel harmonic and percussive component separation by joint modeling of spatial and spectral continuity," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 205–208.

[40] A. Liutkus, D. Fitzgerald, and Z. Rafii, "Scalable audio separation with light kernel additive modelling," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 76–80.

[41] S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard, "Alpha-stable multichannel audio source separation," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[42] P. Magron, R. Badeau, and A. Liutkus, "Lévy NMF for robust nonnegative source separation," in *Proc. IEEE Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[43] H. T. T. Duong, N. Q. K. Duong, Q. C. Nguyen, and C. P. Nguyen, "Multichannel audio source separation exploiting NMF-based generic source spectral model in Gaussian modeling framework," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LCA/ICA)*, 2018.

[44] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010.

[45] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[46] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparse criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 – 1074, 2007.

[47] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito non-negative matrix factorization with group sparsity," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[48] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Proc. Interspeech*, 2012, pp. 17–20.

[49] H. Kuttruff, *Room Acoustics*, 4th ed. New York: Spon Press, 2000.

[50] 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, 2006.

[51] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2012.

[52] N. Ono, Z. Koldovsk, S. Miyabe, and N. Ito, "The 2013 Signal Separation Evaluation Campaign," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2013, pp. 1–6.

[53] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," in *Latent Variable Analysis and Signal Separation (LVAICA)*. Springer, 2015, vol. 9237, pp. 387–395.
[54] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *Trans. Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 549–557, 2011.
[55] N. Bryan and G. Mysore, "An efficient posterior regularized latent variable model for interactive sound source separation," in *Proc. The 30th International Conference on Machine Learning (ICML)*, 2013, pp. 208–216.
[56] Z. Rafii and B. Pardo, "Online REPET-SIM for real-time speech enhancement," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 848–852.
[57] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, pp. 3238–3242.
[58] S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind Speech Separation and Enhancement With GCC-NMF," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 745–755, 2017.
[59] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.

Thanh T. H. Duong received the B.S. degree in Computer Science from Hanoi National University of Education (HNUE), Vietnam, in 2000, and the M.S. degree in Computer Science from Hanoi University of Science and Technology (HUST), Vietnam, in 2008. She is currently a senior lecturer at Hanoi University of Mining and Geology (HUMG), Vietnam. She has been also a Ph.D. student in Computer Science at MICA Institute (HUST) since 2014. Her research focuses on audio source separation and enhancement.

Ngoc Q. K. Duong received the B.S. degree from Posts and Telecommunications Institute of Technology (PTIT), Vietnam, in 2004, and the M.S. degree in Electronic Engineering from Paichai University, Korea, in 2008. He obtained the Ph.D. degree at the French National Institute for Research in Computer Science and Control (INRIA), Rennes, France in 2011.

From 2004 to 2006, he was with Visco JSC as a System Engineer. He was also a Research Engineer for the acoustic echo/noise cancelation system at Emersys Company, Korea in 2008. He is currently a Senior Scientist at Technicolor R&D France where he has worked since Nov. 2011. His research interest concerns signal processing and machine learning, applied to audio, image, and video. He has received several research awards, including the IEEE Signal Processing Society Young Author Best Paper Award in 2012 and the Bretagne Young Researcher Award in 2015. He is the co-author of more than 45 scientific papers and about 30 patent submissions.

Phuong Cong Nguyen received the B.S. degree from Hanoi University of Science and Technology (HUST), Vietnam, in 1999, and the M.S. degree in instrumentation and control systems, the Ph.D. degree in signal processing, both from HUST in 2001 and 2009, respectively. He has been senior lecturer at School of Electrical Engineering (HUST) since 1999 and permanent researcher at MICA Institute (HUST) since 2011. His research interest consists of signal processing and machine learning applied to audio.

Cuong Quoc Nguyen received the engineer (1996), M.S. (1998) degrees in electrical engineering from Hanoi University of Science and Technology(HUST), Vietnam, and PhD degree in Signal-Image-Speech-Telecoms from INP Grenoble (Institut National Polytechnique de Grenoble), France, in 2002. He is a lecturer/researcher at Department of Instrumentation and Industrial Informatics - School of Electrical Engineering - HUST. His research interest concerns Signal Processing, Speech Recognition, Embedded Systems and RF communication.