# Multichannel audio source separation exploiting NMF-based generic source spectral model in Gaussian modeling framework

Thanh Thi Hien Duong, Ngoc Q. K. Duong, Cong-Phuong Nguyen, Quoc-Cuong Nguyen

# Multichannel audio source separation exploiting NMF-based generic source spectral model in Gaussian modeling framework

Thanh Thi Hien Duong[1,2], Ngoc Q. K. Duong[3], Cong-Phuong Nguyen[1,4], and Quoc-Cuong Nguyen[4]

[1]International Research Institute MICA, Hanoi University of Science and Technology, Vietnam,
[2]Information Technology Faculty, Hanoi University of Mining and Geology, Vietnam
[3]Imaging Science Lab, Technicolor, France
[4]Dept. of Instrumentation and Industrial Informatic, Hanoi University of Science and Technology, Vietnam
Email: [1]duongthihienthanh@humg.edu.vn, [3]quang-khanh-ngoc.duong@technicolor.com
[4]{phuong.nguyencong, cuong.nguyenquoc}@hust.edu.vn

**Abstract.** Nonnegative matrix factorization (NMF) has been well-known as a powerful spectral model for audio signals. Existing work, including ours, has investigated the use of generic source spectral models (GSSM) based on NMF for single-channel audio source separation and shown its efficiency in different settings. This paper extends the work to multichannel case where the GSSM is combined with the source spatial covariance model within a unified Gaussian modeling framework. Specially, unlike a conventional combination where the estimated variances of each source are further constrained by NMF separately, we propose to constrain the total variances of all sources altogether and found a better separation performance. We present the expectation-maximization (EM) algorithm for the parameter estimation. We demonstrate the effectiveness of the proposed approach by using a benchmark dataset provided within the 2016 Signal Separation Evaluation Campaign.

**Keywords:** Multichannel audio source separation, generic spectral model, nonnegative matrix factorization, spatial covariance model, Gaussian modeling.

## 1 INTRODUCTION

Audio source separation, which aims at separating individual sound sources from their mixture, is crucial in many practical applications such as speech enhancement, sound post-production, and robotics. Despite numerous efforts in the past decades, its performance in real-world conditions is still far from perfect [1]. To improve the separation performance, depending on specific scenario where certain side information can be known, a range of *informed* source separation algorithms has been proposed in the literature [2]. Such side information can be *e.g.,* score associated with musical sources [3], text associated with spoken speeches [4], motion associated with audio-visual objects in a video [5], or deformed references [6]. Following this trend, very abstract semantic information just about the type of audio source (*e.g.,* if a source in the mixture is speech, musical instrument, or environmental sound) has been used to create a so-called universal speech model in [7] or the universal sound class model in [8]. Exploiting this idea, we have investigated the use of generic speech and noise model for single-channel speech separation in [9] and shown its promising result. Further more, we have proposed to combine the block sparsity constraint investigated in [7] with the component sparsity constraint presented in [8] in a common formulation so as to take into account the advantage of both of them.

It is interesting to note that most cited work above [3–5, 7, 9, 8] considered only a single channel case, where the mixtures are mono, and exploited non-negative matrix factorization (NMF) [10, 11] to model the spectral characteristics of audio sources. When more recording channels are available, multichannel source separation algorithm should be considered as it allows to exploit important information about the spatial locations of the sources. Such additional information has been shown to greatly improve the separation performance. To date, the spatial cues can be modeled by *e.g.,* the interchannel time difference and interchannel intensity difference [12], the rank-1 mixing vector in the frequency domain [13, 14], or the full-rank spatial covariance matrix in Gaussian modeling framework [15, 16]. In this paper, we present an extension of our previous work [9] to multichannel case where the NMF-based GSSM is combined with the powerful full-rank spatial covariance model in a Gaussian modeling paradigm [15]. Note that the combination of NMF with such spatial covariance model has been investigated in several works [17, 18, 16]. However, our work is different from [17, 18] in the sense that we use the pre-trained GSSM so as the intermediate source variances are better constrained. As consequence, the overall algorithm is much less sensitive to the parameter initialization and it does not suffer from the well-known permutation problem. Our work is also different from [16] as we exploit the mixed group sparsity constraint in the optimization algorithm in order to automatically select the most representative spectral components in the GSSM. Specially, unlike all existing approaches [17, 18, 16] where the estimated variances of each source are independently constrained by NMF, we propose to constrain the total variances of all sources altogether so as the parameters are estimated in a more global consistent way.

The structure of the rest of the paper is as follows. We introduce the problem formulation and modeling in Section 2. We then present the proposed multichannel algorithm with the details of parameter estimation in Section 3. We validate the effectiveness of the proposed approach in Section 4. Finally we conclude in Section 5.

## 2 PROBLEM FORMULATION AND MODELING

Let us denote by $\mathbf{c}_j(t) \in \mathbb{R}^{I\times 1}$ the contribution of $j$-th source, $j = 1, 2, ..., J$, to an array of $I$ microphones, and $\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t)$ the mixture signal. The objective of source separation is to estimate $\mathbf{c}_j(t)$ given $\mathbf{x}(t)$. As most source separation algorithm operates in the frequency domain, we denote by $\mathbf{c}_j(n, f)$ and $\mathbf{x}(n, f)$ the short-term Fourier transform (STFT) of $\mathbf{c}_j(t)$ and $\mathbf{x}(t)$, respectively, where $n = 1, 2, .., N$ presents time frame index and $f = 1, 2, ..., F$ the frequency bin index. The mixing model in the frequency domain writes:

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

### 2.1 General Gaussian modeling framework

We consider the nonstationary Gaussian modeling framework [15], where $\mathbf{c}_j(n,f)$ is modeled as a zero-mean complex Gaussian random vector $\mathbf{c}_j(n,f) \sim \mathcal{N}_c(\mathbf{0}, \mathbf{\Sigma}_j(n,f))$. Here $\mathbf{0}$ denotes a $I \times 1$ vector of zeros, and the covariance matrix $\mathbf{\Sigma}_j(n,f)$ is factorized as

$$\mathbf{\Sigma}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f), \tag{2}$$

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. 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

$$\mathbf{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f). \tag{3}$$

Denoting by $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f) = \mathbb{E}(\mathbf{x}(n,f)\mathbf{x}^H(n,f))$ the empirical covariance matrix, which can be numerically computed by local averaging over neighborhoods of $(n,f)$ [15, 16]. The negative log-likelihood is computed as

$$\mathcal{L}(\theta) = \sum_{n,f} \mathrm{tr}\big(\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\widehat{\mathbf{\Sigma}}_{\mathbf{x}}(n,f)\big) + \log\det\big(\pi\mathbf{\Sigma}_{\mathbf{x}}(n,f)\big), \tag{4}$$

where $\det()$ presents the matrix determinant. Under this model, the parameters $\{v_j(n,f), \mathbf{R}_j(f)\}_{j,n,f}$ can be estimated in the Maximum likelihood (ML) sense by minimizing $\mathcal{L}(\theta)$. Then the STFT coefficients of the source images are obtained in the minimum mean square error (MMSE) sense by multichannel Wiener filtering as

$$\hat{\mathbf{c}}_j(n,f) = v_j(n,f)\,\mathbf{R}_j(f)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f). \tag{5}$$

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

### 2.2 NMF-based spectral model

As mentioned earlier, NMF has been widely applied to single channel audio source separation where the spectrogam of the mixture is factorized by two smaller matrices known as the spectral dictionary and the activation [11]. When adapting NMF to the considered Gaussian modeling framework, the nonnegative source variance $v_j(n,f)$ can be approximated as

$$v_j(n,f) \approx \hat{v}_j(n,f) = \sum_{k=1}^{K_j} w_{jfk}h_{jkn}, \tag{6}$$

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 model the $j$-th source. Given the matrix of the source variances $\mathbf{V}_j = \{v_j(n,f)\}_{n,f} \in \mathbb{R}_+^{F\times N}$, the corresponding NMF parameters can be estimated by minimizing the Itakura-Saito divergence, which offers scale invariant property, as

$$\underset{\mathbf{H}_j\geq 0,\mathbf{W}_j\geq 0}{min} D(\mathbf{V}_j\|\mathbf{W}_j\mathbf{H}_j), \tag{7}$$

where $D(\mathbf{V}_j\|\mathbf{W}_j\mathbf{H}_j) = \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}\big(v_j(n,f)\|\hat{v}_j(n,f)\big)$, and $d_{IS}(x\|y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$.

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 [10, 11]. To our best knowledge, this NMF formulation for the source variances within the presenting Gaussian modeling framework was first presented in [17], and then further discussed in [18].

## 3 PROPOSED APPROACH

We will first introduce the GSSM construction in Section 3.1. We then discuss the novel GSSM fitting with a sparsity constraint in Section 3.2. Finally, we present the derived EM algorithm in Section 3.3. *Note that we focus on NMF as spectral model in this paper, however the whole idea of the proposed approach can actually be used for other spectral models than NMF.*

### 3.1 GSSM construction

We assume that the types of sources in the mixture are known and some examples of them are available. This is actually feasible in practice as we often know at least what type of target signal to extract from a recording, *e.g.,* in the speech enhancement usecase, one target source is speech and another is noise. Let the spectrogram of $p$-th example of the $j$-th source $s_j^p(t)$ be denoted by $\mathbf{V}_j^p$. First, $\mathbf{V}_j^p$ is used to learn a corresponding NMF spectral dictionary, denoted by $\mathbf{W}_j^p$, by optimizing the criterion similarly to (7):

$$\underset{\mathbf{H}_j^p\geq 0,\mathbf{W}_j^p\geq 0}{min} D(\mathbf{V}_j^p\|\mathbf{W}_j^p\mathbf{H}_j^p) \tag{8}$$

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

$$\mathbf{U}_j = [\mathbf{W}_j^1, \ldots, \mathbf{W}_j^{P_j}], \tag{9}$$

then the GSSM for all the sources is computed by

$$\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_J]. \tag{10}$$

As an example for speech and noise separation, in practical implementation we may need several speech examples from different male voices and female voices (*e.g.,* $P_1 = 4$), and several examples of different types of noise such as those from outdoor environment, cafeteria, waterfall, street (*e.g.,* $P_2 = 5$).

### 3.2 GSSM fitting with mixed group sparsity constraint

The GSSM for all sources $\mathbf{U}$ constructed in (10) become a large matrix when the number of examples $P_j$ for each source increases, and it is actually a redundant dictionary since different examples may share similar spectral patterns. Thus in the NMF model fitting, sparsity constraint is naturally needed so as to automatically select only a subset of $\mathbf{U}$ which represents the sources in the mixture [19]. In other words, the model-based spectrogram of the mixture $\widetilde{\mathbf{V}} = \sum_{j=1}^{J} \mathbf{V}_j$ is decomposed by solving the following optimization problem

$$\underset{\mathbf{H} \geq 0}{min}\; D(\widetilde{\mathbf{V}} \| \mathbf{U}\mathbf{H}) + \lambda \Omega(\mathbf{H}) \tag{11}$$

where $\Omega(\mathbf{H})$ presents a penalty function imposing sparsity on the activation matrix $\mathbf{H} \in \mathbb{R}_{+}^{K \times N}$, and $\lambda$ is a trade-off parameter determining the contribution of the penalty. Note that unlike existing approaches [17, 18, 16] where the matrix of the estimated variances of each source $\mathbf{V}_j$ was constrained by NMF independently as (7), we propose here to constrain the matrix of the total variances of all sources $\widetilde{\mathbf{V}}$ altogether by (11). This can be seen as an additional NMF-based separation step applied on the source variances, while the existing works does not perform any additional separation of the variances, but more like denoising of the already separated variances. In our recent work [9] we investigated a general form for the penalty function as

$$\Omega(\mathbf{H}) = \alpha \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_g\|_1) + (1-\alpha) \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1), \tag{12}$$

where the first term on the right hand side of the equation presents the so-called *block* sparsity-inducing penalty (which enforces the activation of relevant examples only while omitting irrelevant ones since their corresponding activation block in $\mathbf{H}$ will likely converge to zero), the second term presents the so-called *component* sparsity-inducing penalty (which enforces the activation of relevant components in $\mathbf{U}$ only), $\alpha \in [0, 1]$ weights the contribution of each term. In (12), $\mathbf{h}_k \in \mathbb{R}_{+}^{1 \times N}$ is a row (or component) of $\mathbf{H}$, $\mathbf{H}_g$ is a subset of $\mathbf{H}$ representing the activation coefficients for $g$-th block, $G$ is the total number of blocks, $\epsilon$ is a non-zero constant (*i.e.,* set by $3 * 10^{-6}$ in our experiment), and $\|.\|_1$ denotes $\ell_1$-norm operator (*i.e.,* the maximum absolute column sum of the matrix). In the considered setting, a block represents one training example for a source and $G$ is the total number of used examples (*i.e.,* $G = \sum_{j=1}^{J} P_j$).

By putting (12) into (11), we have a complete criterion for estimating $\mathbf{H}$ given $\widetilde{\mathbf{V}}$ and the pre-trained spectral model $\mathbf{U}$. The derived MU rule for updating $\mathbf{H}$ is presented in [9] and summarized in the Algorithm 1, where $\mathbf{Y}_g$ is a uniform matrix of the same size as $\mathbf{H}_g$ and $\mathbf{z}_k$ a uniform row vector of the same size as $\mathbf{h}_k$.

### 3.3 Proposed multichannel algorithm

Within the presenting Gaussian modeling framework, EM algorithm has been derived 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*. In the E-step, the Wierner filter $\mathbf{Q}_j(n,f)$ and the expected covatiance $\widehat{\boldsymbol{\Sigma}}_j(n,f)$ of the spatial images of the $j$-th source are computed. Then in the M-step, $\mathbf{R}_j(f)$ and $v_j(n,f)$ are updated by minimizing (4), which gives close-form solution. The detail of this EM derivation can be found in [15, 18]. For the proposed approach as far as the GSSM concerned, the E-step of EM algorithm remains the same. In the M-step, we additionally perform the optimization defined in (11) by MU rules so as the estimated intermediate source variance $v_j(n,f)$ is further updated with the supervision of the GSSM. The detail of EM algorithm for the parameter estimation is summarized in Algorithm 1.

**Algorithm 1** EM algorithm for the parameter update

```
// E-step (perform calculation for all $j$, $n$, $f$):
$\boldsymbol{\Sigma}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)$ // equation (2)
$\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\,\mathbf{R}_j(f)$ // equation (3)
$\mathbf{Q}_j(n,f) = \boldsymbol{\Sigma}_j(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)$
$\widehat{\boldsymbol{\Sigma}}_j(n,f) = \mathbf{Q}_j(n,f)\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{Q}_j^H(n,f) + \big(\mathbf{I} - \mathbf{Q}_j(n,f)\big)\boldsymbol{\Sigma}_j(n,f)$

// M-step (perform calculation for all $j$, $n$, $f$)
$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{v_j(n,f)}\widehat{\boldsymbol{\Sigma}}_j(n,f)$ // update $\mathbf{R}_j(f)$
$v_j(n,f) = \frac{1}{I}\text{tr}(\mathbf{R}_j^{-1}(f)\widehat{\boldsymbol{\Sigma}}_j(n,f))$ // update $v_j(n,f)$

$\mathbf{V}_j = \{v_j(n,f)\}_{n,f}$
$\widetilde{\mathbf{V}} = \sum_{j=1}^{J} \mathbf{V}_j$
// Perform NMF in the M-step to further constrain source spectra by the GSSM
for $iter = 1, ..., $ MU-iteration do
    for $g = 1, ..., G$ do
        $\mathbf{Y}_g \leftarrow \frac{1}{\epsilon + \|\mathbf{H}_g\|_1}$
    end for
    $\mathbf{Y} = [\mathbf{Y}_1^T, \ldots, \mathbf{Y}_G^T]^T$
    for $k = 1, ..., K$ do
        $\mathbf{z}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_k\|_1}$
    end for
    $\mathbf{Z} = [\mathbf{z}_1^T, \ldots, \mathbf{z}_K^T]^T$
    $\widehat{\mathbf{V}} = \mathbf{UH}$
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{U}^T(\widetilde{\mathbf{V}} \odot \widehat{\mathbf{V}}^{\cdot -2})}{\mathbf{U}^T(\widehat{\mathbf{V}}^{\cdot -1}) + \lambda(\alpha\mathbf{Y} + (1-\alpha)\mathbf{Z})}\right)^{\cdot \frac{1}{2}}$ // MU rule
end for

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

# 4 EXPERIMENTS

## 4.1 Dataset and settings

We validated the performance of the proposed algorithm in a popular but very important speech enhancement usecase where we know already two types of sources in the mixture: speech and noise. For 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 [1]. This devset contained stereo mixtures of 10 second duration and 16 KHz sampling rate. They were mixed of male/female speeches and noises recorded from six different public environments: cafeteria (Ca), square (Sq), and subway (Su). Overall there were nine mixtures of two sources: three with Ca noise, four with Sq noise, and two with Su noise. The signal-to-noise ratio was drawn randomly per mixture between -17 and +12 dB by the dataset creators.

For training the GSSM for speech and noise, we took one male voice and two female voices from the SiSEC 2015[2]. These three speech examples were also 10-second long. Five noise training examples were extracted from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND)[3]. Again they were 10-second long and contained three types of environmental noise: cafeteria, square, metro. We made sure that these examples used for GSSM training are different from those in the devset, which were used for testing. The number of NMF components in $\mathbf{W}_j^p$ for each speech example was set to 32, while that for noise example was 16, and each $\mathbf{W}_j^p$ was obtained after 20 MU iterations. Other parameter settings were as follows. The STFT window length of 50% overllaping was 1024. The spatial covariance matrix $\mathbf{R}_j(f)$ for noise was initialized following the diffuse model, while $\mathbf{R}_j(f)$ for speech was initialized following the direct+diffuse model [15] assuming the direction-of-arrival (DoA) for speech source is $90^0$. For testing, we first varied the number of EM and MU iterations and found that generally the convergence obtained after about 10 iterations. Specifically the best result was obtained by 15 EM iterations and 10 MU iterations. The trade-off parameter $\lambda$ determining the contribution of the sparsity-inducing penalty in (11) and the factor $\alpha$ weighting the contribution of each penalty term in (12) were tested with different values: $\lambda = \{1, 10, 25, 50, 100, 200, 500\}$, $\alpha = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and we found that the algorithm is less sensitive to the choice of $\alpha$, while more sensitive to the choice of $\lambda$ and $\lambda > 10$ decreases the separation performance. The best choice for these parameters are $\lambda = 10, \alpha = 0.2$.

### 4.2 Comparison results

We compare the speech separation performance of the proposed approach with several state of the art and baseline algorithms as follows:

- Liu's method: the algorithm performed Time Difference of Arrival (TDOA) clustering based on GCC-PHAT and participated to the same SiSEC 2016 campaign [1]. The separation results were submitted by the authors and evaluated by the SiSEC organizers.
- Wood's method [20]: this algorithm first applied NMF to the magnitude spectrograms of the mixtures with channels concatenated in time. Each dictionary atom

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

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

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

was then clustered to either the speech or the noise according to its spatial origin. Again the separation results for devset were submitted to the SiSEC 2016 campaign and evaluated by the SiSEC organizers.

- Arberet's method [17]: using the similar Gaussian modeling framework, the algorithm further constrained the estimated source variances by unsupervised NMF where the parameters were obtained by optimizing the criterion (7) in the M-step of EM algorithm instead of (11) like us. Such optimization criterion was implemented by Ozerov *et. al.* in [18].
- Baseline 1: the presenting GSSM + full-rank spatial covariance approach but there is no sparsity constraint in (11) (*i.e.,* $\lambda = 0$). This is to investigate the importance of the sparsity constraint (12) in the GSSM fitting.
- Baseline 2: the presenting GSSM + full-rank spatial covariance approach but the estimated variances of each source $\mathbf{V}_j$ are further constrained by NMF where the corresponding activation matrix $\mathbf{H}_j$ obtained by optimizing the following criterion:

$$\underset{\mathbf{H}_j \geq 0}{min}\ D(\mathbf{V}_j \| \mathbf{U}_j \mathbf{H}_j) + \lambda \Omega(\mathbf{H}_j) \tag{13}$$

  We submitted results obtained by this method to the SiSEC 2016 BGN task and obtained the best performance among other submitting methods in term of the overall signal-to-distortion (SDR) ratio [1].
- Proposed method: the presenting GSSM + full-rank spatial covariance approach where the matrix of the total variances of all sources $\widetilde{\mathbf{V}}$ is constrained by NMF and the activation matrix is obtained by optimizing (11). EM algorithm for the corresponding parameter updates is present in Algorithm 1.

The separation performance (for speech source only) for all approaches was evaluated by 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 [21]. These values are shown in Table 1 where the higher the better.

| Methods | SDR | SIR | SAR | ISR |
|---|---|---|---|---|
| Liu's method | -7.0 | -1.4 | **15.0** | 3.1 |
| Wood's method [20] | 1.9 | 3.6 | 3.7 | 5.1 |
| Arberet's method [17, 18] | 4.4 | 4.6 | 12.1 | **15.9** |
| Baseline 1 (No sparsity constraint) | 0.4 | -1.1 | 9.5 | 8.3 |
| Baseline 2 ($\lambda = 10, \alpha = 0.2$) | 7.4 | 8.9 | 12.7 | 11.3 |
| Proposed method ($\lambda = 10, \alpha = 0.2$) | **7.7** | **10.7** | 11.6 | 13.9 |

**Table 1.** Average speech separation performance obtained on the devset of the BGN task of SiSEC 2016. Results for Liu's method and Wood's method were submitted by the authors [1].

It is interesting to see that the result obtained by the Baseline 1 is lower than that of Arberet's method, even the former used the pre-trained GSSM while the later was 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 unsupervised NMF case where the spectral patterns were randomly initialized and then updated by MU rules. The importance of such sparsity penalty is explicitly confirmed by the fact that the result obtained by the Baseline 2 was far more better than that of the Baseline 1. It is also not surprising to see that the Baseline 2 clearly outperforms Arberet's method as the former exploited additional information about the types of sources in the mixtures so as to learn the GSSM in advance. We also tested the case where the small size dictionary obtained by jointly decomposing all training examples for the target signal, but the performance was lower than the Baseline 2. Finally, the proposed method offers the best separation performance in terms of SDR and SIR, the two important criteria. This confirms the effectiveness of the proposed approach where the GSSM is successfully combined with the spatial covariance model in a unified Gaussian modeling framework. Furthermore, the benefit of the new criterion (11) compared to the conventional one (13) for the NMF parameter estimation is supported. *Our further analysis, which is not described here due to the lack of space, shows in addition that with such new criterion, the algorithm is less sensitive to the parameter initialization and the choice of hyper-parameters $\lambda$ and $\alpha$ as compared to the Baseline 2.*

## 5 CONCLUSION

We have presented a novel multichannel audio source separation algorithm, which exploits the use of generic source spectral model within the well-established Gaussian modeling framework. Such redundant GSSM can be easily learned from source examples by NMF and shown to be very useful in guiding the source separation. Specially, we have proposed a new optimization criterion in order to better constrain the intermediate source variances estimated in each EM iteration. Experiment with a benchmark dataset from the SiSEC 2016 campaign has confirmed the effectiveness of the proposed approach compared to both the state of the art and the baselines. Motivating by the GSSM, future work can be devoted to extend the current approach so as to exploit in addition the use of a *generic spatial covariance model*, which remains to be defined.

## References

1. 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.
2. A. Liutkus, J. L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS).* IEEE, 2013, pp. 1–4.
3. 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.

4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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 SoICT*, 2015, pp. 247–251.
10. 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.
11. 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.
12. M. Mandel and D. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 275–278.
13. 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.
14. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 276–280.
15. 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.
16. 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.
17. 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 ISSPA*, 2010, pp. 1–4.
18. 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.
19. A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito non-negative matrix factorization with group sparsity," in *Proc. IEEE ICASSP*, 2011, pp. 21–24.
20. S. Wood and J. Rouat, "Blind speech separation with GCC-NMF," in *Proc. Interspeech*, 2016, pp. 3329–3333.
21. 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.