



**HAL**  
open science

# A Tractable Framework for Estimating and Combining Spectral Source Models for Audio Source Separation

Simon Arberet, Alexey Ozerov, Frédéric Bimbot, Rémi Gribonval

► **To cite this version:**

Simon Arberet, Alexey Ozerov, Frédéric Bimbot, Rémi Gribonval. A Tractable Framework for Estimating and Combining Spectral Source Models for Audio Source Separation. [Research Report] RR-7556, 2011, pp.27. inria-00572249v1

**HAL Id: inria-00572249**

**<https://inria.hal.science/inria-00572249v1>**

Submitted on 1 Mar 2011 (v1), last revised 4 May 2012 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

*A Tractable Framework for Estimating and  
Combining Spectral Source Models  
for Audio Source Separation*

Simon Arberet — Alexey Ozerov — Frédéric Bimbot — Rémi Gribonval

N° 7556

Mars 2011

— Audio, Speech, and Language Processing —

 *rapport  
de recherche*

ISRN INRIA/RR--7556--FR+ENG

ISSN 0249-6399



## A Tractable Framework for Estimating and Combining Spectral Source Models for Audio Source Separation

Simon Arberet<sup>\*</sup>, Alexey Ozerov<sup>†</sup>, Frédéric Bimbot<sup>‡</sup>, Rémi  
Gribonval<sup>§</sup>

Theme : Audio, Speech, and Language Processing  
Perception, Cognition, Interaction  
Équipes-Projets METISS

Rapport de recherche n° 7556 — Mars 2011 — 27 pages

**Abstract:** The underdetermined blind audio source separation (BSS) problem is often addressed in the time-frequency (TF) domain assuming that each TF point is modeled as an independent random variable with sparse distribution. On the other hand, methods based on structured spectral model, such as the Spectral Gaussian Scale Mixture Models (Spectral-GSMMs) or Spectral Non-negative Matrix Factorization models, perform better because they exploit the statistical diversity of audio source spectrograms, thus allowing to go beyond the simple sparsity assumption. However, in the case of discrete state-based models, such as Spectral-GSMMs, learning the models from the mixture can be computationally very expensive. One of the main problem is that using a classical Expectation-Maximization procedure often leads to an exponential complexity with respect to the number of sources. In this paper, we propose a framework with a linear complexity to learn spectral source models (including discrete state-based models) from noisy source estimates. Moreover, this framework allows combining probabilistic models of different nature that can be seen as a sort of probabilistic fusion. We illustrate that methods based on this framework can significantly improve the BSS performance compared to the state-of-the-art approaches.

**Key-words:** Blind source separation, multichannel audio, Gaussian mixture model, expectation-maximization algorithm, convolutive mixture

This work was supported in part by the EU FET-Open project FP7-ICT-225913-SMALL and the OSEO, the French State agency for innovation, under the Quaero program.

<sup>\*</sup> Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland. e-mail: simon.arberet@epfl.ch

<sup>†</sup> alexey.ozarov@inria.fr

<sup>‡</sup> frederic.bimbot@irisa.fr

<sup>§</sup> remi.gribonval@inria.fr

# Un cadre de Complexité Maîtrisable pour Estimer et Combiner des Modèles Spectraux pour la Séparation de Sources Audio

**Résumé :** La séparation aveugle de sources audio (SAS) est souvent traitée dans le plan temps-fréquence (TF), en partant de l'hypothèse que chaque point TF est la réalisation d'une variable aléatoire indépendante ayant une distribution parcimonieuse. D'autre part, les méthodes fondées sur un modèle spectral, telles que les modèles de mélanges de gaussiennes spectrales (MMG-Spectraux) ou les modèles de factorisation en matrices non-négatives spectrales (FMN-spectraux), obtiennent de meilleurs résultats parce qu'ils exploitent la diversité statistique des spectrogrammes des sources audio, permettant ainsi d'aller au-delà de la simple hypothèse de parcimonie. Cependant, dans le cas des modèles à états discrets, tels que les MMG-Spectraux, l'apprentissage à partir du mélange peut être d'une complexité rédhibitoire. Un des problèmes majeurs est que l'utilisation de la procédure Espérance-Maximisation (EM) aboutit à une complexité calculatoire exponentielle par rapport au nombre de sources. Dans cet article, nous proposons un cadre, d'une complexité calculatoire linéaire, pour apprendre des modèles de sources (y compris des modèles à états discrets) à partir d'estimations bruitées des sources. De plus, ce cadre permet de combiner des modèles probabilistes de différentes natures et permet ainsi de faire une sorte de "fusion" probabiliste. Nous montrons que des méthodes construites à partir de ce cadre permettent d'améliorer les performances de SAS par rapport aux méthodes de l'état de l'art.

**Mots-clés :** Séparation aveugle de sources, audio multicanal, modèles de mélanges de sources, Espérance-Maximisation, mélanges convolutifs

## 1 Introduction

Most audio recordings can be viewed as mixtures of several audio signals (e.g., musical instruments or speech), called *source signals* or *sources*, that are usually active simultaneously. The sources may have been mixed synthetically with a mixing console or by recording a real audio scene using microphones.

The mixing of  $N$  audio sources on  $M$  channels is often formulated as the following convolutive mixing model:

$$x_m(\tau) = \sum_{n=1}^N \sum_{l=0}^{L-1} a_{mn}(l) s_n(\tau - l), \quad 1 \leq m \leq M, \quad (1)$$

where  $s_n(\tau)$  and  $x_m(\tau)$  denote sampled time signals of respectively the  $n$ -th source and the  $m$ -th mixture ( $\tau$  being a discrete time index), and  $a_{mn}(l) \in \mathbb{R}$  denote the finite (sampled) impulse response of some causal filter.

The goal of the *convolutive Blind Source Separation (BSS)* problem is to estimate the  $N$  source signals  $s_n(\tau)$  ( $1 \leq n \leq N$ ), given the  $M$  mixture signals  $x_m(\tau)$  ( $1 \leq m \leq M$ ).

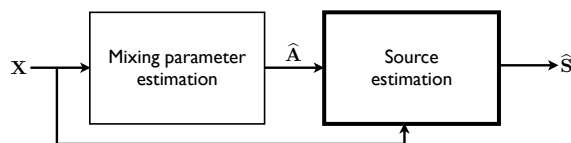


Figure 1: Block diagram of the two step approach. In this paper, we focus on the second block i.e. the source estimation.

When the number of sources is larger than the number of mixture channels ( $N > M$ ), the BSS problem is said to be *underdetermined* and is often addressed by sparsity-based approaches [1, 2, 3, 4] consisting in the following two steps:

- first the mixing parameters are estimated at the first step, and
- usually, the second step, consisting in source coefficients estimation, is solved with the minimum mean squared error (MMSE) estimator given a sparse source prior and the mixing parameters.

Since audio signals are usually not sparse in the time domain, the estimation of the source coefficients is done in some time-frequency (TF) domain by using for example the short time Fourier transform (STFT). Figure 1 shows the block diagram of the two step approach in the STFT domain, where the mixing equation (1) is usually approximated as follows [5]<sup>1</sup>:

$$\mathbf{X}(t, f) \approx \mathbf{A}(f) \mathbf{S}(t, f), \quad (2)$$

where  $\mathbf{S}(t, f) = [S_1(t, f), \dots, S_N(t, f)]^T$  and  $\mathbf{X}(t, f) = [X_1(t, f), \dots, X_M(t, f)]^T$  ( $t = 1, \dots, T$  and  $f = 1, \dots, F$  being time and frequency indices) denote respectively column vectors of source and mixture STFTs; and  $\mathbf{A}(f) = [A_{mn}(f)]_{m,n=1}^{M,N}$

<sup>1</sup>In this paper we adopt the following generic notations. Time and frequency indices are always noted in parentheses. Lower case letters are used for time domain quantities and upper case letters for their STFTs. Vectors and matrices with respect to dimensions  $M$  and  $N$  are denoted using bold letters.

is an  $M \times N$  complex-valued mixing matrix with elements  $A_{mn}(f)$  being the discrete Fourier transforms of filters  $\mathbf{a}_n(l) = [a_{1n}(l), \dots, a_{Mn}(l)]^T$ .

Sparse prior distributions that are mostly used for audio sources include Laplacian [6, 2], generalized Gaussian [7], Student-t [8], and mixtures of two Gaussians [9]. One of the main limitations of the sparsity-based two-step approaches [1, 2, 3, 4] is that each TF coefficient is assumed to be independent of the others. As a consequence, the redundancy and structure of each audio source are not taken into account.

This issue has been partially addressed by the Local Gaussian Model (LGM) [10], where source TF coefficients are locally modeled by Gaussian distributions with free variances. However, this method exploits only a neighborhood of each TF point, in order to estimate the parameters of the corresponding Gaussian distribution.

A more globally structured approach consists in assuming a structured spectral model of each source, such as Spectral Gaussian Mixture Model (Spectral-GMM) [11, 12], the Spectral Gaussian Scaled Mixture Model (Spectral-GSMM) [13] or the Spectral Nonnegative Matrix Factorization (Spectral-NMF) model [14]. The Spectral-GMM and Spectral-GSMM have been successfully used to separate sources in the single channel case ( $M = 1$ ) [11, 12, 13], where sparsity-based methods become unsuitable. However, this approach cannot be considered as blind because the models need to be learned from some training sources which are supposed to have characteristics very close to those of the sources to be separated. An Expectation-Maximization (EM) algorithm can, in principle, be used to learn spectral models directly from the mixture [15], but this approach suffers from the following issues:

- *Computationally intractable inference:* In the case of state-based models such as GMM and GSMM the number of Gaussian components in the observation density grows exponentially with the number of sources, which often leads to a computationally intractable exact inference.
- *Sensitivity to initialization:* The algorithm is very sensitive to the initialization, i.e., it can converge to an unsatisfactory local maximum depending on the initial values of parameters. This second issue also concerns the Spectral-NMF model.

## 1.1 Contributions

We propose an approach which enables the learning of discrete state-based models with a tractable complexity that is linear with respect to (w.r.t.) the number of sources. Our approach is inspired by the idea that one can first separate sources with some method, providing source estimates  $\tilde{S}_n$ , and then learn the Spectral models from these estimates. This way, each source model can be learned separately, which results in a tractable linear computational complexity w.r.t. the number of sources.

However, a potential danger of such an approach is that the resulting models, learned from the source estimates instead of the true sources, could be altered by the errors contained in the estimated sources.

### 1.1.1 Learning in linear time from noisy observations

So as to take into account these errors, we develop a general framework to learn source models from noisy observations. The idea consists in modeling the errors produced by the first algorithm as a noise, estimating the parameters of this noise (e.g., variances if we assume a Gaussian noise) with a moment matching approach [16], and then learning the source models from the source estimates  $\tilde{S}_n$  while taking into account the noise.

This framework allows to reduce the computational complexity from exponential to linear in the number of sources. It also makes it possible to take into account the errors produced at the first separation layer.

### 1.1.2 Multilayer “model fusion”

Once the source models are learned, it is possible to add one more *learning layer* by running the previous steps with the same or with a different source models. At the end, the sources can be estimated from the mixture with the MMSE estimator given the models computed within the last layer.

Moreover, this framework allows the combination of probabilistic models of different nature (e.g., here we combine LGM with Spectral-GMM, Spectral-GSMM and Spectral-NMF models) in a *multi-layer* fashion that can be seen as a sort of probabilistic fusion. For example, such a multi-layer approach can be useful in the following configuration: for each source, one can learn a model (on the first layer) with an estimation algorithm that would be robust to the initialization conditions, and then, learn a model (on the next layer) which represents better the source characteristics.

Finally, the proposed framework is quite general. It can be applied to various probabilistic models and is not limited to source separation. For example, it is in line with the noisy speech recognition framework by Deng *et al* [17], where a Gaussian model is used to model the error of clean speech features estimation from noisy speech. In summary this framework offers:

- a computationally tractable approximate inference in factorial state-based models; and
- the ability to combine probabilistic models of different nature in a multi-layer fashion, while preventing error propagation from one layer to another, which can be useful, as explained, to overcome the issue of sensibility to initialization.

## 1.2 Related works

The proposed framework is related to that of learning GMMs from incomplete data by Ghahramani and Jordan [18], and extends it in the sense that [18] becomes a partial case of our framework when all noise variances are very small or very large. Moreover, we detail our framework for other models (i.e., GSMM and NMF).

In the context of source separation, the idea to combine different methods in a multi-layer fashion relates to [19], where single channel singing voice separation problem is processed in two steps. First, a pitch-based inference method is applied to estimate singing voice location in the TF domain. Second, a music



background model (NMF model) is learned from incomplete data (i.e., ignoring the regions where singing voice was detected) as in [18]. As compared to [19], our framework is more general and investigated in this paper in the case of multi-channel source separation.

The proposed framework is also related to the variational approach by Attias [20, 21] where on each iteration of the EM algorithm the sources are assumed independent conditionally to the observations and the posterior of each source is approximated with a GMM. In this paper, the posterior of the sources are also assumed independent but the approach is more modular in the sense that it is not restricted to only one source model: it is possible to specify a different source model for each layer and, in line with [22]<sup>2</sup>, for each source.

It should also be noted that our framework, as detailed in this paper, implements several existing source models (i.e., GMM, GSMM and NMF) in the same setting, thus allowing a fair experimental comparison between these models. Moreover, the framework brings a common point of view on several machine learning approaches, such as learning from incomplete data [18] and variational inference in factorial state-based models in the single [23] and multi-channel [20] cases.

### 1.3 Working assumptions

We assume that the mixing is underdetermined and either instantaneous or convolutive. In this paper, we are interested in source models estimation (see the second block “source estimation” of Fig. 1), and we assume that the mixing parameters are known.

### 1.4 Organization of the paper

The paper is organized as follows. In section 2, we describe the state-of-the-art source models (LGM, Spectral-GMM, Spectral-GSMM and Spectral-NMF model) that can be incorporated into our framework. In section 3, a high level presentation of our framework, based on learning spectral models from noisy observations, is given. In section 4, we explain how to model the errors of the source estimates for all source models introduced in section 2. We then derive algorithms for learning Spectral-GSMM and Spectral-NMF models from noisy observations. Finally, in section 6, we evaluate the performance of our approach on mixtures of monophonic, polyphonic and percussive music sources and compare it to the state-of-the-art approaches. Preliminary aspects of this work were presented in [24] in the case of Spectral-GMM source model and linear instantaneous mixtures. In this paper we extend it to convolutive mixtures and to other Spectral models (Spectral-GSMM and Spectral-NMF) thus providing a more consistent experimental evaluation.

## 2 Source models

In this section, we present state-of-the-art source models that can be incorporated in the proposed framework.

---

<sup>2</sup>Note that in [22] there is no combination of models in a multi-layer fashion.

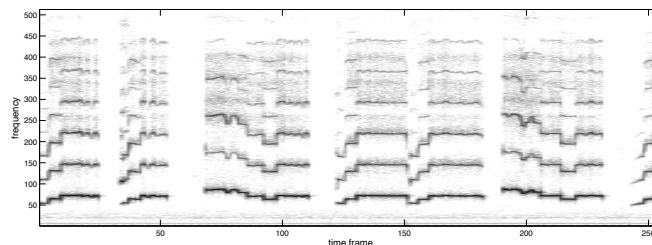


Figure 2: Spectrogram of a flute performance. The amplitude of each TF point is represented as shades of gray varying from black for the strongest amplitude to white for the weakest.

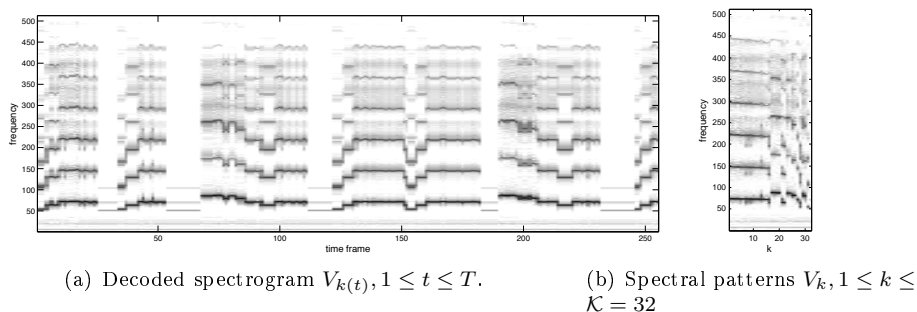


Figure 3: Decoded spectrogram and spectral patterns of a Spectral-GMM learned from the flute signal depicted in Figure 2.

As discussed in the introduction, most of audio sources have a structure in the TF domain. An example spectrogram of a flute performance is represented on Figure 2. Many frames in this spectrogram are very similar, which suggests that it can be well represented by a small number of characteristic spectral patterns. This motivates the use of structure spectral models which we describe hereafter.

LGM [10], Spectral-GMM [11, 12], Spectral-GSMM [11, 13] and Spectral-NMF [14] are all zero-mean Gaussian<sup>3</sup> source models, but with different assumptions on the structure of the source variance in the TF domain. While the LGM model assumes the variance to be free in each TF point, the Spectral-GMM, Spectral-GSMM and the Spectral-NMF condition the constraint structure source variances by representing the source short-time spectra with a limited number  $\mathcal{K}$  of characteristic spectral patterns. In the next subsections, we describe these four models and explain how they are related to each other. For the sake of simplicity, the source index  $n$  is dropped for a while.

<sup>3</sup> Probability density function (pdf) of Spectral-GMM is in fact a sum of Gaussian pdfs, but conditionally to a given state sequence this model is Gaussian. In the same way, Spectral-GSMM is Gaussian conditionally to a given state sequence and a particular sequence of scaling coefficients.

## 2.1 Local Gaussian Model (LGM)

In the LGM model [10], the source TF coefficients  $S(t, f)$  are assumed to be realizations of independent zero-mean complex-valued Gaussian variables with variances  $\sigma^2(t, f)$ :

$$p(S(t, f) | \sigma^2(t, f)) = N_c(S(t, f); 0, \sigma^2(t, f)), \quad (3)$$

where  $N_c(V; \mu, \Sigma)$  is the probability density function (pdf) of a circular complex Gaussian vector  $S \in \mathbb{C}^p$  expressed as:

$$N_c(S; \mu, \Sigma) \triangleq \pi^{-p} [\det(\Sigma)]^{-1} \exp \left[ - (S - \mu)^H \Sigma^{-1} (S - \mu) \right] \quad (4)$$

where “ $(\cdot)^H$ ” denotes conjugate-transpose,  $\mu$  is the  $p$ -dimensional complex-valued mean vector and  $\Sigma$  is the  $p \times p$  complex-valued covariance matrix. The variance  $\sigma^2(t, f)$  can be estimated in the neighborhood of the corresponding TF point using some symmetrical overlapping two-dimensional window [10].

## 2.2 Spectral Gaussian Mixture Model (GMM)

We define the short time Fourier spectrum  $S(t) \triangleq [S(t, f)]_f$  as a column vector composed of the elements  $S(t, f)$ ,  $f = 1, \dots, F$  of source  $S$  at time frame  $t$ . In the Spectral-GMM approach, the short time Fourier spectrum  $S(t)$  of each source is modeled as a multidimensional zero-mean complex valued GMM with pdf given by:

$$p(S(t) | \lambda^{gmm}) = \sum_{k=1}^{\mathcal{K}} \pi_k N_c(S(t); \bar{0}, \Sigma_k), \quad (5)$$

where  $\bar{0}$  is a vector of all zeros,  $\pi_k$  (satisfying  $\sum_{k=1}^{\mathcal{K}} \pi_k = 1$ ) and  $\Sigma_k$  denote respectively the weight and the diagonal covariance matrix of the  $k$ -th GMM state. Introducing nonnegative column vectors of variances  $V_k \triangleq [\sigma_k^2(f)]_f$ ,  $k$ -th statecovariance matrix can be expressed as  $\Sigma_k = \text{diag}(V_k)$ . The set of all parameters of the Spectral-GMM of source  $S$  is denoted as  $\lambda^{gmm} \triangleq \{\pi_k, \Sigma_k\}_k$ .

This source model can be viewed as a two-step generative process, where for each frame  $t$ , the first step is to pick one state  $k(t)$  with the corresponding characteristic spectral pattern  $V_{k(t)} \in \mathbb{V} = \{V_1, \dots, V_{\mathcal{K}}\}$  and the corresponding probability  $P(q(t) = k(t)) = \pi_{k(t)}$ , where  $q(t)$  denotes the random variable of the state at frame  $t$ . Second, given the state  $k(t)$  picked at frame  $t$ , every source TF coefficient  $S(t, f)$  is independently generated from a zero-mean Gaussian distribution with variance  $\sigma_{k(t)}^2(f)$ :

$$p(S(t, f) | \lambda^{gmm}, k(t)) = N_c \left( S(t, f); 0, \sigma_{k(t)}^2(f) \right). \quad (6)$$

One can interpret (6) as the pdf of a Gaussian model with parameters  $\theta^{gmm} \triangleq \{\lambda^{gmm}, \{k(t)\}_t\}$ . We use this interpretation later in section 3.2 and we call  $\theta^{gmm}$  the *complete* model of the Spectral-GMM  $\lambda^{gmm}$ .

As opposed to the LGM model, where there are as much free parameters (variances) as TF points, the Spectral-GMM is defined by  $\mathcal{K} \times F$  free parameters (variances) only. Figure 3(b) depicts the spectral patterns of a 32-states GMM

learned from the flute signal represented in Figure 2 and Figure 3(a) represents the decoded sequence of spectral patterns which looks quite similar as the original spectrogram of Figure 2. As opposed to the LGM, where the variances of each TF points are estimated independently, the Spectral-GMM exploits a global structure of the signal. As a consequence, less free parameters are to be estimated and thus their estimation is likely to be statistically more consistent.

However, GMM does not model explicitly amplitude variation of sounds with the same spectral shape. As a result, different components might be used to represent the same spectral pattern with different amplitude levels, which may lead to less consistent estimation of model parameters. To overcome this issue the GSMM was proposed [13].

### 2.3 Spectral Gaussian Scaled Mixture Models (GSMM)

The GSMM model [13] is a variant of the GMM which includes a scaling parameter  $g_k(t)$  for each component  $k$  and for each frame  $t$  so that the spectral patterns  $\mathbb{V}$  are invariant to the amplitude variation of frames in the observed signal.

The pdf of the Spectral-GSMM is given by:

$$p(S(t)|\lambda^{gsmm}) = \sum_{k=1}^{\mathcal{K}} \pi_k N_c(S(t); \bar{0}, g_k(t) \Sigma_k). \quad (7)$$

The set of all parameters of the Spectral-GSMM of source  $S$  is denoted as  $\lambda^{gsmm} \triangleq \{\pi_k, \Sigma_k, \{g_k(t)\}_t\}_k$ . As for the GMM, this source model can be viewed as a two-step process where, for each frame  $t$ , the first step is to pick one state  $k(t)$  corresponding to the spectral pattern  $V_{k(t)} \in \mathbb{V} = \{V_1, \dots, V_{\mathcal{K}}\}$  with probability  $P(q(t) = k(t)) = \pi_{k(t)}$ . The distribution of the TF source coefficients given the state  $k(t)$  at frame  $t$  is a zero-mean Gaussian with variance  $g_k(t)\sigma_{k(t)}^2(f)$ :

$$p(S(t, f)|\lambda^{gsmm}, k(t)) = N_c\left(S(t, f); 0, g_k(t)\sigma_{k(t)}^2(f)\right). \quad (8)$$

We also consider the *complete* Spectral-GSMM model  $\theta^{gsmm} \triangleq \{\lambda^{gsmm}, \{k(t)\}_t\}$  defined by:

$$p(S(t, f)|\theta^{gsmm}) = N_c\left(S(t, f); 0, g_k(t)\sigma_{k(t)}^2(f)\right). \quad (9)$$

Note that, as opposed to the Spectral-GMM model, the variances of the sources  $V(t) = [\sigma^2(t, f)]_{f=1}^F = g_k(t)V_k$  can be any scaled version of one of the spectral pattern of the set  $\mathbb{V}$ .

### 2.4 Spectral Nonnegative Matrix Factorization (NMF)

In the Spectral-NMF model, the short time Fourier spectrum  $S(t)$  of each source is modeled by as follows:

$$p(S(t, f)|\lambda^{nmf}) = N_c\left(S(t, f); 0, \sum_{k=1}^{\mathcal{K}} h_k(t)v_k(f)\right), \quad (10)$$

where  $V = [v_k(f)]_{f,k} = [V_1, \dots, V_{\mathcal{K}}]$  and  $H = [h_k(t)]_{k,t}$  are two non-negative matrices [14]. The set of the Spectral-NMF parameters is denoted by  $\lambda^{nmf} =$

$\{V, H\}$ . Note that, as opposed to the Spectral-GMM model, the variances of the sources  $V(t) = [\sigma^2(t, f)]_{f=1}^F = \sum_{k=1}^{\mathcal{K}} h_k(t) V_k$  are not picked from a discrete set  $\mathbb{V}$  of characteristic spectral patterns, but belong to the convex cone  $\mathbb{V}_c$  generated by  $\mathbb{V}$ .

### 3 General formulation of the framework

In this section, we present a framework to learn spectral source models from noisy measurements with a linear complexity w.r.t. the number of sources. We first explain why learning Spectral-GSMM source models from the mixture data with an EM algorithm is untractable as the number of sources increases. We then present our approach in section 3.2.

#### 3.1 Learning from the mixture

Given the Spectral-GSMM<sup>4</sup> source models  $\mathbf{\Lambda}^{\text{gsmm}} = [\lambda_n^{\text{gsmm}}]_{n=1}^N$  introduced in section 2, the density of the mixture  $\mathbf{X}(t, f)$  according to the mixing model (2) is a multichannel Spectral-GSMM with  $\mathcal{K}^N$  states [20]. In order to learn the Spectral-GSMM using an EM algorithm directly from the mixture  $\mathbf{X}(t, f)$ , it is necessary to compute the posterior probabilities (sometimes called *responsibilities* [20, 25]) that  $\mathbf{k} \triangleq [k_n]_{n=1}^N$  is the active state at frame  $t$ :

$$\gamma_{\mathbf{k}}(t) \triangleq P(\mathbf{q}(t) = \mathbf{k} | \mathbf{X}(t); \mathbf{A}(f), \mathbf{\Lambda}^{\text{gsmm}}) \quad (11)$$

where  $\mathbf{q}(t) \triangleq [q_n(t)]_{n=1}^N$ .

In order not to have to compute the  $\mathcal{K}^N$  responsibilities  $\gamma_{\mathbf{k}}(t)$  at each iteration of the EM algorithm, we propose an alternative approach in the following section where each source model is learned separately. As a consequence the algorithmic complexity of our approach will be linear (instead of exponential) with respect to the number of sources, and thus tractable.

#### 3.2 Proposed Approach

The proposed framework is composed of three steps which are described in this section.

**Step 1: moment estimation** The proposed framework (depicted in Figure 4) assumes the existence of a BSS method, based on a source model  $\lambda^{\text{init}}$ , that can also provide for each source and each TF point  $(t, f)$  the following generalized posterior moments (we drop the source index and the TF indices for sake of legibility):

$$m_1 = \mathbb{E} \{ S | \mathbf{X}, \lambda^{\text{init}} \} \quad (12)$$

$$m'_{2k} = \mathbb{E} \{ (S - m_1)^k (S - m_1)^{*k} | \mathbf{X}, \lambda^{\text{init}} \}. \quad (13)$$

Note that, if a random variable is circular, the moments and cumulants with a different number  $k$  of conjugate terms and non-conjugate terms are all

<sup>4</sup>Spectral-GMM is a special case of Spectral-GSMM with all the gains fixed to one.

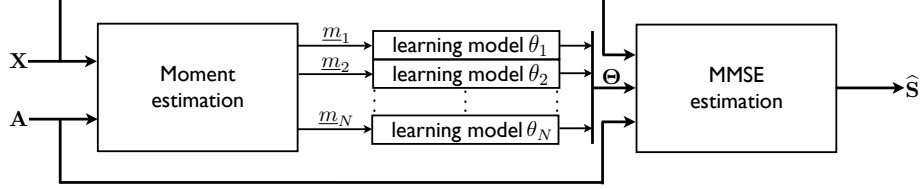


Figure 4: Block diagram of the proposed BSS framework.  $\mathbf{X}$  is the mixture,  $\mathbf{A}$  is the mixing matrix,  $\underline{m}_n = \{m_{1,n}, m'_{2k,n}\}$  are the posterior moments of source  $n$  defined in (12) and (13),  $\Theta = [\theta_n]_{n=1}^N$  is the concatenation of the *complete* source model parameters and  $\hat{\mathbf{S}}$  is the final estimation of the sources.

zero [26]. We assume that the random variable  $E \triangleq S - m_1$  given  $\mathbf{X}$  is circular. As a consequence, we only consider the centered moments  $m'$  having the same number  $k$  of conjugate terms and non-conjugate terms.

**Step 2: model inference** For each source we consider the so-called decoupled noisy model:

$$\tilde{S} = S + E, \quad (14)$$

where  $\tilde{S}$  is the source estimation given by the MMSE estimator, that is  $m_1$ , and  $E$  is a random variable (noise) with moments  $m'_{2k}$ . The noise  $E$  models the error in the source estimation  $\tilde{S}$  by the initial BSS method based on the source model  $\lambda^{init}$ . Note that, if we consider a circular Gaussian model for the noise  $E$ , we only need to estimate  $m'_2$  which corresponds to the noise variance  $\sigma_e^2$ . On the other hand, the higher moments  $m'_{2k}, k > 1$  can be used if we want to consider circular non-Gaussian noise.

Now we consider the inference of a source model  $\lambda$  (e.g.  $\lambda = \lambda^{gsmm}$ ) from the decoupled noisy model(14). This problem can be interpreted as a single sensor denoising problem, i.e., a source separation problem [11,12] with observed mixture  $\tilde{S}$  and two latent sources: *target source*  $S$  and *noise*  $E$ . The model is learned by optimizing the ML criterion:

$$\lambda = \arg \max_{\lambda'} p(\tilde{S} | \lambda', \lambda_e), \quad (15)$$

where  $\lambda_e$  is the noise model which parameters are estimated by matching them with the moments (13).

In the case of discrete state-based models like GMM and GSMM, we also estimate the most likely states given by:

$$\begin{aligned} k(t) &= \arg \max_{k'} P(q(t) = k' | \tilde{S}, \lambda, \lambda_e) \\ &= \arg \max_{k'} \gamma_{k'}(t). \end{aligned} \quad (16)$$

As mentioned in sections 2.2 and 2.3, given the state-decoded sequence  $\{k(t)\}_t$  and the source model parameters  $\lambda$ , the source can be assumed to be distributed according to a simpler model (a Gaussian distribution if  $\lambda$  is a GMM or a GSMM) with parameters  $\theta = \{\lambda, \{k(t)\}_t\}$ .

So as to somehow unify state-based models with non state-based models, we define the *complete* model  $\theta$  as:

$$\theta := \begin{cases} \{\lambda, \{k(t)\}_t\} & \text{in the case where } \lambda \text{ is a state-based model} \\ & \text{with a state decoding sequence } \{k(t)\}_t, \\ \{\lambda\} & \text{otherwise.} \end{cases}$$

As a consequence,  $\theta$  denotes a model that can be considered as none state-based, and thus its MMSE source estimates is not combinatorial. In the case of the GMM, the MMSE of the complete model  $\theta^{gmm}$  is sometime called the ‘‘hard’’ estimator while the MMSE of model  $\lambda^{gmm}$  is called the ‘‘soft’’ estimator [27].

We derive, in section 5, EM algorithms to optimize the ML criteria (15) and (16) in the case of Spectral-GSMM and Spectral-NMF source models.

**Step 3: source separation** Once the source models have been learned, the sources can be separated from the mixture data (2) and the source model parameters  $\Theta = [\theta_n]_{n=1}^N$  with the MMSE estimator.

Note that, in this framework, the source models are learned separately, which leads to a linear complexity algorithm (instead of exponential complexity). Note also that, as long as we can provide the moments (12) and (13), we can run the framework with one model  $\lambda$  and then run the framework again with a possibly different model. We will explain in section 4 how the posterior moments (12) and (13) can be computed in the case of the source models presented in section 2.

## 4 Posterior moments

The latent sources  $\mathbf{S}(t, f) = [S_n(t, f)]_{n=1}^N$  are assumed to be realizations of random variables following a known probability distribution with parameters  $\Theta$ .

Let assume that, given the model  $\Theta$  and the mixture coefficients  $\mathbf{X}(t, f)$ , the source coefficients  $\mathbf{S}(t, f)$  are independent from  $\{\mathbf{X}(t', f')\}_{(t', f') \neq (t, f)}$ , so that

$$p(\mathbf{S}(t, f) | \mathbf{X}(t, f), \Theta) = p(\mathbf{S}(t, f) | \{\mathbf{X}(t, f)\}_{t, f}, \Theta). \quad (17)$$

If  $\Theta := \mathbf{A} = [\lambda_n]_{n=1}^N$ , then this independence assumption holds for the LGM and NMF models but not for the GMM and GSMM models because the states  $\mathbf{q}(t)$  of each frame  $t$  depend of the values of  $\mathbf{X}(t, f), \forall f$  according to (11). However if we consider the complete source models  $\Theta := \{\mathbf{A}, \mathbf{\Gamma}\}$ , i.e. if the state decoding sequence  $\mathbf{\Gamma} = [\mathbf{k}(t)]_{t=1}^T$  is included into the set of model parameters  $\Theta$  as suggested in Step 2 of section 3.2, then the independence assumption (17) holds for these models.

Then, given this posterior distribution, the estimator  $\delta(\cdot)$  that minimizes the posterior risk  $\mathbf{r}(t, f) = \mathbb{E}\{L(\mathbf{S}(t, f), \delta(\mathbf{X}(t, f))) | \mathbf{X}(t, f), \Theta\}$ , with  $L(\cdot, \cdot)$  being the squared error loss  $L(\mathbf{S}, \delta(\mathbf{X})) = \|\mathbf{S} - \delta(\mathbf{X})\|^2$ , i.e. the MMSE estimator, is the Bayes estimator  $\delta_{\Theta}(\mathbf{X}(t, f)) = \mathbb{E}\{\mathbf{S}(t, f) | \mathbf{X}(t, f), \Theta\}$ .

For convenience, let us associate the scalar moments (12) and (13) to the column vectors  $\mathbf{m}_1(t, f)$  and  $\mathbf{m}_2'(t, f)$  that respectively stack the N sources moments  $m_1(t, f)$  and  $m_2'(t, f)$ .

Thus, the vectors  $\mathbf{m}_1(t, f)$  and  $\mathbf{m}'_2(t, f)$ , required by the proposed framework of section 3.2 are:

$$\mathbf{m}_1(t, f) := \delta_{\Theta}(\mathbf{X}(t, f)) \quad (18)$$

$$\mathbf{m}'_2(t, f) := \text{diag}([\mathcal{R}_{n,n}(t, f)]_n). \quad (19)$$

where  $\mathcal{R}(t, f) = \mathbb{E} \left\{ [\mathbf{S}(t, f) - \delta(\mathbf{X}(t, f))] [\mathbf{S}(t, f) - \delta(\mathbf{X}(t, f))]^H \middle| \mathbf{X}(t, f), \Theta \right\}$  is the posterior covariance matrix.

In the case of the zero mean source Gaussian prior  $p(\mathbf{S}(t, f) | \Theta) = N_c(\mathbf{S}(t, f); \mathbf{0}, \Sigma(t, f))$ , the Bayes estimator is:

$$\delta_{\Theta}(\mathbf{X}(t, f)) = \mathbf{W}(t, f) \mathbf{X}(t, f) \quad (20)$$

where  $\mathbf{W}(t, f)$  is the Wiener gain computed as follows:

$$\mathbf{W}(t, f) \triangleq \Sigma(t, f) \mathbf{A}^H(f) (\mathbf{A}(f) \Sigma(t, f) \mathbf{A}^H(f))^{-1}. \quad (21)$$

The resulting posterior covariance matrix is:

$$\mathcal{R}(t, f) = [\mathbf{I} - \mathbf{W}(t, f) \mathbf{A}(f)] \Sigma(t, f). \quad (22)$$

Note that in the case of the Gaussian source prior, the posterior is also Gaussian, and thus all higher-order cumulants are equal to zero and the posterior moments  $\mathbf{m}'_{2k}(t, f), k > 1$  are just given by known formulas depending only on  $\mathbf{m}'_2(t, f)$ . Thus, in the case of the Gaussian source prior, there is no need to estimate other moments than  $\mathbf{m}_1(t, f)$  and  $\mathbf{m}'_2(t, f)$ .

As mentioned in section 2, LGM, the *complete*<sup>5</sup> Spectral-GSMM and Spectral-NMF are all Gaussian source models, but with different assumptions on the structure of the source variances in the TF domain. Thus, the posterior moments of these models are given by Eqs. (18) - (22). The only difference between these models is the structure on the *a priori* source covariance matrix  $\Sigma(t, f) = \mathbb{E}[\mathbf{S}(t, f) \mathbf{S}(t, f)^H | \Theta]$  which is diagonal since the sources are assumed independent. The diagonal entries of  $\Sigma(t, f)$  for the models LGM, Spectral-GSMM and Spectral-NMF are defined in section 2.

## 5 Model inference via the EM algorithm

In section 5.1, we derive an EM algorithm in the case where the source model is a Spectral-GSMM ( $\lambda := \lambda^{gsmm}$ ) and in section 5.2 we derive an EM algorithm in the case where the source model is a Spectral-NMF ( $\lambda := \lambda^{nmf}$ ).

In both cases, we consider a circular Gaussian noise  $E(t) = [E(t, f)]_f$  in the decoupled noisy model(14), i.e.  $p(E(t) | \lambda_e) = N_c(E(t); \bar{0}, \Sigma_e(t))$ , with  $\Sigma_e(t) = \text{diag}([\sigma_e^2(t, f)]_f)$ . Then we only have to match  $m'_2(t, f)$  to the noise variances  $\sigma_e^2(t, f) := m'_2(t, f)$ .

### 5.1 Learning Spectral-GSMMs from noisy observations

Algorithm 1 summarizes an EM algorithm optimizing the ML criterion (15). Mathematical derivation of this algorithm is very similar to [12].

<sup>5</sup>see its definition in section 2.2



Diagonals of covariance matrices  $\Sigma_k$  can be initialized by a K-means clustering algorithm applied to the source estimate  $\tilde{S}$ .

The Spectral-GMM model is a special case of the Spectral-GSMM with gains equal to one. Therefore the EM algorithm for Spectral-GMM is the same as Algorithm 1 without step 4 and with  $g_k^{(l)}(t) = 1, \forall k, t, l$ .

It can easily be checked that algorithm 1 is more general than the EM algorithm for learning GMMs from incomplete data [18] in the sense that if a subset of noise variances tends to zero (observed TF points), and the rest of the variances to  $+\infty$  (missing TF points), Algorithm 1 and the EM algorithm from [18] become similar.

---

**Algorithm 1** EM Algorithm for source Spectral-GSMM estimation in the ML sense (index  $(l)$  in power denotes the parameters estimated at the  $l^{th}$  iteration of the algorithm) and  $\lambda = \lambda^{gsmm}$

---

1. Compute the weights  $\gamma_k^{(l)}(t)$  satisfying  $\sum_k \gamma_k^{(l)}(t) = 1$  and

$$\gamma_k^{(l)}(t) \triangleq P(q(t) = k | \tilde{S}, \lambda^{(l)}, \lambda_e) \propto \pi_k^{(l)} N_c(\tilde{S}(t); \bar{0}, g_k^{(l)}(t) \Sigma_k^{(l)} + \Sigma_e(t)) \quad (23)$$

where  $q(t)$  is the current state of GSMM  $\lambda$  at frame  $t$ .

2. Compute the expected Power Spectral Density (PSD) for state  $k$

$$\begin{aligned} \langle |S(t, f)|^2 \rangle_k^{(l)} &\triangleq \mathbb{E}_S \left[ |S(t, f)|^2 \mid q(t) = k, \tilde{S}, \lambda^{(l)}, \lambda_e \right] = \\ &= \frac{g_k^{(l)}(t) \sigma_k^{2,(l)}(f) \sigma_e^2(t, f)}{g_k^{(l)}(t) \sigma_k^{2,(l)}(f) + \sigma_e^2(t, f)} + \left| \frac{g_k^{(l)}(t) \sigma_k^{2,(l)}(f) \cdot \tilde{S}(t, f)}{g_k^{(l)}(t) \sigma_k^{2,(l)}(f) + \sigma_e^2(t, f)} \right|^2 \end{aligned}$$

3. Re-estimate Gaussian weights

$$\pi_k^{(l+1)} = \frac{1}{T} \sum_t \gamma_k^{(l)}(t)$$

4. Re-estimate Gaussian gains

$$g_k^{(l+1)}(t) = \frac{1}{F} \sum_f \frac{\langle |S(t, f)|^2 \rangle_k^{(l)}}{\sigma_k^{2,(l)}(f)}$$

5. Re-estimate covariance matrices

$$\sigma_k^{2,(l+1)}(f) = \frac{1}{\sum_t \gamma_k^{(l)}(t)} \sum_t \frac{\langle |S(t, f)|^2 \rangle_k^{(l)}}{g_k^{(l+1)}(t)} \gamma_k^{(l)}(t)$$


---

### 5.1.1 Decoding step

Once the source models  $\mathbf{\Lambda}^{gsmm} = [\lambda_n^{gsmm}]_{n=1}^N$  learned, one could estimate the sources from the mixture as explained in section 3. However, in that case, the

decoding step that consists in estimating all the mixture responsibilities  $\gamma_{\mathbf{k}}(t)$  of (11) at each frame  $t$ , is of complexity  $O(\mathcal{K}^N)$ .

In order to avoid decoding with exponential complexity on  $N$ , we compute the responsibilities of each source with (23) by executing step 1) of the last iteration ( $L+1$ ) of algorithm 1. As algorithm 1 is performed independently on each source, the decoding step has a linear complexity on  $N$ :  $O(N\mathcal{K})$ . Also, instead of using the MMSE estimator given the model  $\Lambda^{\text{gsmm}}$ , i.e.  $\delta_{\Lambda^{\text{gsmm}}}(\mathbf{X}(t, f)) = \sum_{\mathbf{k}} \gamma_{\mathbf{k}}(t) \mathbf{W}_{\mathbf{k}}(t, f) \mathbf{X}(t, f)$  which implies again computing a sum over the  $\mathcal{K}^N$  states, we consider the complete model  $\Theta^{\text{gsmm}} := \{\Lambda^{\text{gsmm}}, \Gamma\}$ , with  $\Gamma = [\mathbf{k}^*(t)]_{t=1}^T$  and  $\mathbf{k}^*(t) = [k_n^*(t)]_{n=1}^N$ . That is, we keep only the most likely state given by:

$$\begin{aligned} k_n^*(t) &= \arg \max_k P(q(t) = k | \tilde{S}_n, \lambda_n^{(L+1)}, \lambda_{e,n}) \\ &= \arg \max_k \gamma_{k_n}^{(L+1)}(t). \end{aligned} \quad (24)$$

The source coefficients can then be estimated with the MMSE estimator given  $\Theta^{\text{gsmm}}$ :

$$\delta_{\Theta^{\text{gsmm}}}(\mathbf{X}(t, f)) = \mathbf{W}_{\mathbf{k}^*(t)}(t, f) \mathbf{X}(t, f).$$

## 5.2 Learning Spectral-NMF from noisy observations

Algorithm 2 summarizes an EM algorithm for the optimization of criterion (15) when  $\lambda$  is a Spectral-NMF model. Parameters  $h_k(t)$  and  $v_k(f)$  can be initialized with an NMF decomposition using, e.g., multiplicative update (MU) rules and Kullback-Leibler (KL) divergence as in [14].

## 6 Experimental study

We evaluate the framework described in section 3 using LGM as an initial BSS method ( $\lambda = \lambda^{\text{init}}$ ) and consider the following source models: Spectral-GMM, Spectral-GSMM and Spectral-NMF model. When the framework is run with a single layer, we call these methods respectively *LGM-GMM*<sup>6</sup>, *LGM-GSMM* and *LGM-NMF*. When the framework is run with  $k$  layers, we name these methods by concatenating the model that is learned for each layer. For example, LGM-GMM-GMM is the two-layer method using LGM as an initial BSS method and where Spectral-GMM is learned in the first and the second layer.

In this section we propose experiments aiming at:

- *Experiment 1*: evaluating the performance of the framework with only one layer. Thus we want to compare LGM with LGM-GMM, LGM-GSMM and LGM-NMF;
- *Experiment 2*: evaluating the performance of the framework with two layers. Thus we want to compare LGM and the one layer methods (we only select the one shows the best results, to keep the figure readable) with LGM-GMM-GMM, LGM-GSMM-GSMM and LGM-NMF-NMF;

<sup>6</sup>LGM-GMM method was also entered to the 2008 Signal Separation Evaluation Campaign on instantaneous underdetermined mixtures, and has shown competitive results as compared to the state-of-the-art approaches (see Table 2. of [28]).

---

**Algorithm 2** EM Algorithm for source Spectral-NMF estimation in the ML sense (index  $(l)$  in power denotes the parameters estimated at the  $l^{\text{th}}$  iteration of the algorithm) and  $\lambda = \lambda^{nmf}$

---

1. Compute the expected variance  $\hat{u}_k^{(l)}(t, f)$  of each component (E-step)

$$\hat{u}_k^{(l)}(t, f) = \left( 1 - \frac{h_k^{(l)}(t)v_k^{(l)}(f)}{\sum_k h_k^{(l)}(t)v_k^{(l)}(f) + \sigma_e^2(t, f)} \right) h_k^{(l)}(t)v_k^{(l)}(f) + \left| \frac{h_k^{(l)}(t)v_k^{(l)}(f) \cdot \tilde{S}(t, f)}{\sum_k h_k^{(l)}(t)v_k^{(l)}(f) + \sigma_e^2(t, f)} \right|^2$$

2. Re-estimate the parameters (M-step)

$$v_k^{(l+1)}(f) = \frac{1}{T} \sum_t \frac{\hat{u}_k^{(l)}(t, f)}{h_k^{(l)}(t)}$$

$$h_k^{(l+1)}(t) = \frac{1}{F} \sum_f \frac{\hat{u}_k^{(l)}(t, f)}{v_k^{(l)}(f)}$$

3. Normalize  $V$  and  $H$  as in [14].
- 

- *Experiment 3*: evaluating the performance of the framework with the state-of-the-art. Thus we want to compare our framework with other structured source model based approaches like LGM and Multichannel-NMFover various mixture conditions including instantaneous and convolutive mixtures.

## 6.1 Experimental particularities

### 6.1.1 Experimental material

We evaluate the methods over stereo mixtures ( $M = 2$ ) of music sources, with the number of sources  $N$  varying from 3 to 6. We do not consider the case of  $N = 2$  which corresponds to the determined case, because in this case the Wiener filter simply becomes the inverse of the mixing matrix ( $\mathbf{W}(t, f) = \mathbf{A}^{-1}(f)$ ), and thus the separation does not depend on the estimated source models and is perfect in the instantaneous case assuming  $\mathbf{A}$  is known. For each experiment, 20 mixtures are generated from different source signals (taken randomly from a dataset of music signals) of duration 10 s, sampled at 16 kHz. We test 3 datasets of sources composed respectively of monophonic, polyphonic and drums musical instruments so as to evaluate the performance of the algorithms depending on the source properties.

### 6.1.2 Performance measure

The performance measure is the Signal-to-Distortion Ratio (SDR) defined in [29]. The results in this section are given in SDR gain w.r.t. the classical DUET method [3]. However the absolute source separation performance in SDR of all tested methods are shown in the tables in the annexe.

### 6.1.3 Parameters

The STFT is computed with a sine window of length 2048 (128 ms). The time-frequency neighborhoods of the LGM method is a 3 by 3 rectangular window in the instantaneous case and 10 by 1 rectangular window in the convolutive case. The size of these neighborhoods yielded the best results on our preliminary experiments on respectively instantaneous and convolutive mixtures. The Spectral-GMMs, Spectral-GSMMs and Spectral-NMF models are learned with 40 iterations of, respectively, algorithms 1 and 2. The Multichannel-NMF method is run for 300 iterations using fixed mixing matrices  $\mathbf{A}(f)$ . We fixed the number  $\mathcal{K}$  of states per Spectral-GSMM source model to 8 and a value of  $\mathcal{K} = 8$  components per Spectral-NMF source model, because these values yielded the best results on our preliminary experiments on instantaneous mixtures.

## 6.2 Instantaneous mixtures

For each  $N$ , a mixing matrix is simulated as described in [30], given an angle of  $50 - 5N$  degrees between successive sources.

### 6.2.1 Results

**Experiment 1** Figure 5 compares the average SDR (improvement w.r.t. DUET) achieved by LGM and the proposed LGM- $\{\text{GMM,GSMM,NMF}\}$  methods. The results show that running one layer of the proposed framework with any of the considered models  $\{\text{GMM,GSMM,NMF}\}$  improves the SDR of LGM by several dBs, especially on the monophonic dataset when there are few sources. For instance, there is approximately 3 dBs of improvement w.r.t. LGM and 6 dB of improvement w.r.t. DUET in the case of 3 sources. We also notice that the results are quite similar across the different models.

**Experiment 2** Figure 6 compares the average SDR (improvement w.r.t. DUET) achieved by LGM and single-layer (e.g. LGM-GSMM) and two-layer (e.g. LGM-GSMM-GSMM) methods of our framework. For the sake of legibility, we only plot the results of one of the single-layer method showed the best results within experiment 1, i.e., LGM-GSMM. The results show that running two layers of our framework with any of the proposed GSMM models improves the SDR by several dBs (nearly 3 dBs of improvement on the monophonic dataset) compared to the single-layer implementation. The improvement is more important when the number of sources is less than 5. The two-layer method based on GSMM (i.e. LGM-GSMM-GSMM) performs the best, in most of the configurations, among all the two-layer methods.

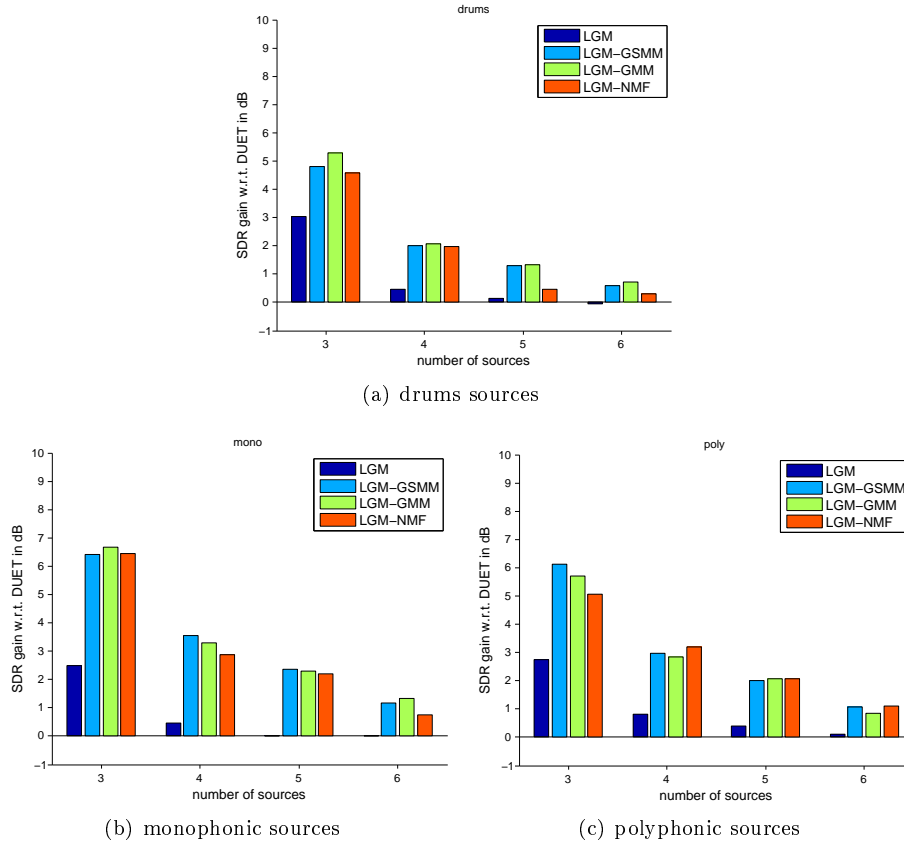


Figure 5: Experiment 1: Source separation gain w.r.t. DUET of LGM and some methods based on our proposed one-layer framework: LGM-GMM, LGM-GSMM and LGM-NMF on instantaneous mixtures with respect to the number of sources.

**Experiment 3** Figure 7 compares the average SDR (improvement w.r.t. DUET) achieved by classical methods like LGM and Multichannel-NMF with LGM initialization (LGM-MNMF) with LGM-GSMM-GSMM and LGM-MNMF-GSMM (i.e. GSMM learned with our framework on the top of LGM-MNMF). We can notice that LGM-MNMF performs better than LGM-GSMM-GSMM on most of the configurations, however LGM-MNMF-GSMM also improves LGM-MNMF on most of the configurations. The performance improvement (w.r.t. DUET) of LGM-MNMF-GSMM in the case of 3 sources, is between 7 dBs and 8 dBs on the drums and polyphonic dataset and nearly 10 dBs on the monophonic dataset. This improvement is more important when there are few sources. These results show the benefit of exploiting the spectral structure of the sources via a collection of spectral shapes.

### 6.3 Convolutional mixtures

Now we compare the same methods as in the previous section on synthetic convolutional mixtures, i.e., static sources filtered by synthetic room impulse responses

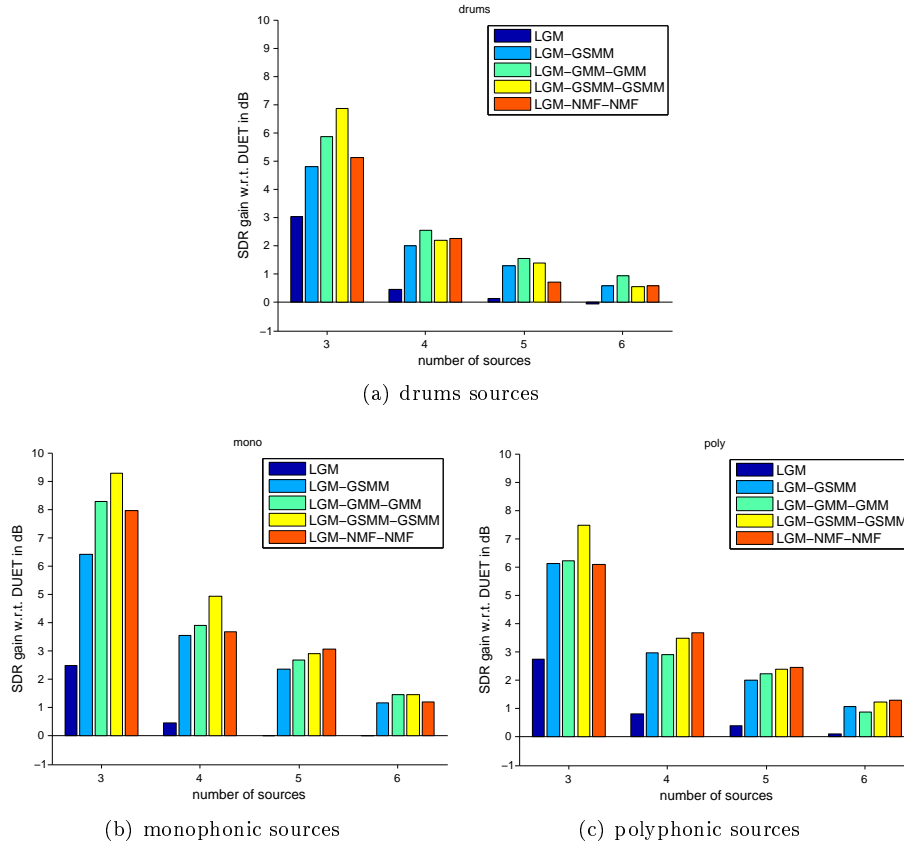


Figure 6: Experiment 2: Source separation gain w.r.t. DUET of LGM and some methods based on our two-layer framework: LGM-GMM-GMM, LGM-GSMM-GSMM and LGM-NMF-NMF on instantaneous mixtures with respect to the number of sources.

simulating a pair of omnidirectional microphones via the Roomsim toolbox [31]. The room dimensions are: 4.45 x 3.55 x 2.5 m. The reverberation time (RT) is set to 130 ms and the distance between the two microphones to 1 m.

### 6.3.1 Estimation of the oracle filters

The estimation of the filters expressed by frequency dependent mixing matrices  $\mathbf{A}(f)$  has been done in the Fourier domain in an *oracle* manner by computing a Principal Component Analysis (PCA) on each frequency band of the source spatial images. The spatial image  $s_{mn}^{img}(\tau)$  of source  $n$  on channel  $m$  is the contribution of this source to the observed mixture in this channel [32], that is:

$$s_{mn}^{img}(\tau) = a_{mn} * s_n(\tau)$$

where  $*$  is the convolution operator and  $a_{mn}$  is the filter of source  $n$  on channel  $m$  defined in the introduction. Let  $S_{mn}^{img}(t, f)$  be the STFT of  $s_{mn}^{img}(\tau)$ , then the source spatial image at time-frequency point  $(t, f)$  is defined by:  $\mathbf{S}_n^{img}(t, f) =$

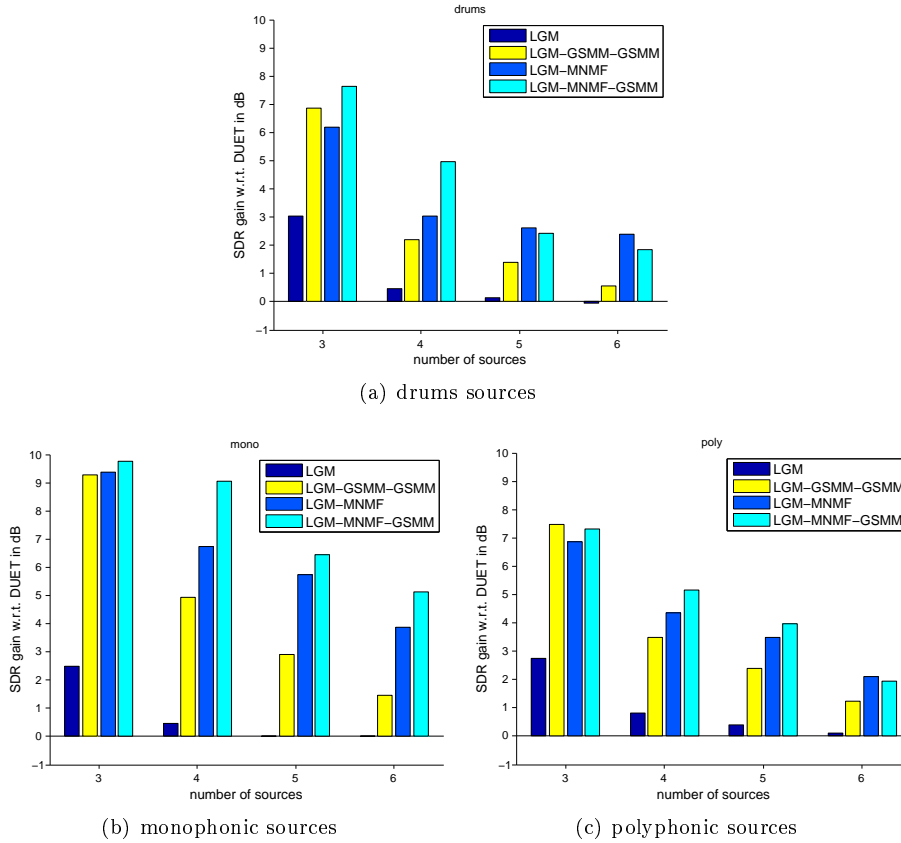


Figure 7: Experiment 3: Source separation gain w.r.t. DUET of state-of-the-art methods (LGM, LGM-MNMF i.e. Multichannel-NMF with LGM initialization) and some methods based on our framework: LGM-GSMM-GSMM, LGM-MNMF-GSMM on instantaneous mixtures with respect to the number of sources.

$[S_{1n}^{img}(t, f), \dots, S_{Mn}^{img}(t, f)]^T$ . Our estimation of  $\mathbf{A}_n(f)$  is given by the principal component of the PCA processed on the data set  $\{S_n^{img}(t, f)\}_t$ .

Note that the DUET [3] and the Multichannel NMF [33] also address the mixing matrices estimation. Here, as we evaluate only the source estimation task, we consider that  $\mathbf{A}(f)$  is the same for all the tested methods and is given by the above mentioned PCA procedure.

### 6.3.2 Results

The results on convolutive mixtures are depicted in Figure 8 and Tables 2. Unsurprisingly, the absolute performance of all the tested methods are worth than in the instantaneous case, because of the narrowband approximation of (2) and because the problem is intrinsically more difficult. However the methods based on a structured spectral models (LGM-MNMF, LGM-MNMF-GSMM, LGM-MNMF-GSMM-GSMM) improve the SDR by several dBs. Adding one or two layers of GSMM to LGM-MNMF increases slightly (up to 1 dB depend-

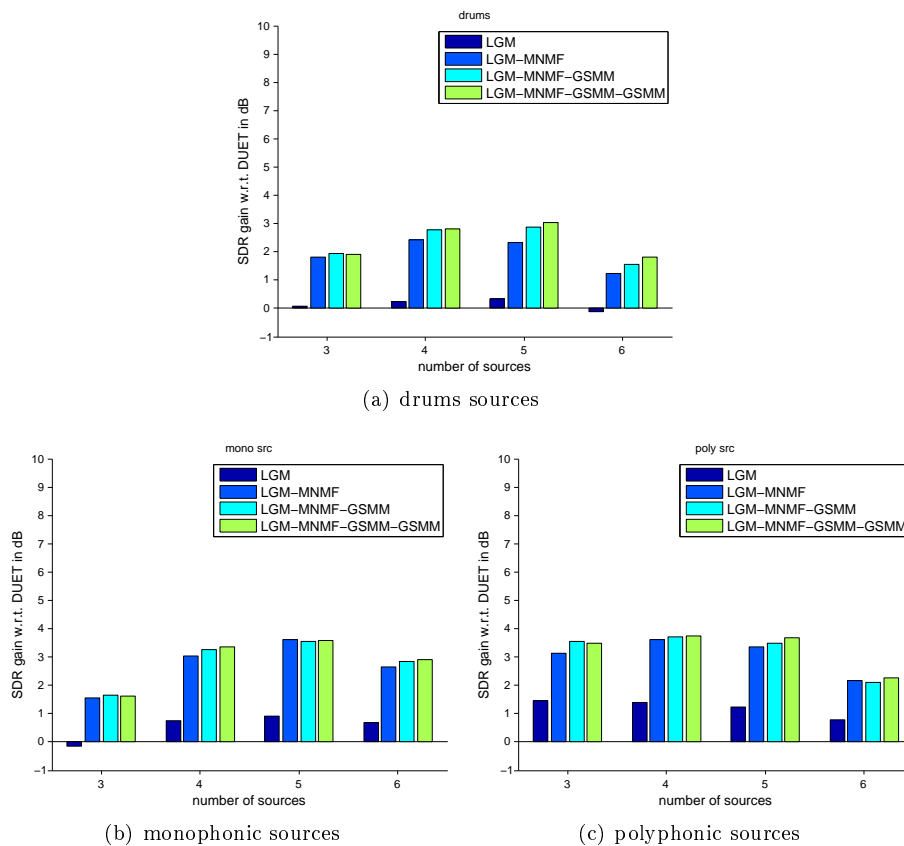


Figure 8: Source separation gain w.r.t. DUET of state-of-the-art methods (LGM, LGM-MNMF i.e. Multichannel-NMF with LGM initialization) and some methods based on our framework: LGM-MNMF-GSMM, LGM-MNMF-GSMM-GSMM on convolutive mixtures with respect to the number of sources.

ing on the configuration) the performance. It is also interesting to notice that, as opposed to the instantaneous case, the improvement is not more important when there are few sources.

## 7 Summary and Conclusions

We have presented a new framework to represent audio sources with spectral source models for multichannel and potentially underdetermined convolutive mixtures. As an initialization the proposed framework requires that a first estimation of the sources in the time-frequency domain is provided as well as an estimation of the posterior moments. The proposed inference method is based on a EM algorithm which runs on each source independently. The resulting complexity of this algorithm grows linearly with the number of sources, and thus remains tractable for real-world mixtures with any number of sources. Moreover, the proposed framework is multi-layer and can combine probabilistic models of different nature that can be seen as a sort of probabilistic fusion. It is



for example possible to learn different source models on the same source while taking into account the error of the source estimate.

We have evaluated our learning framework with LGM and the state-of-the-art Multichannel-NMF for the initialization and with different spectral source models including Spectral-GSMM and Spectral-NMF comparing them with the state-of-the-art methods on stereo underdetermined mixtures in various settings. These settings include instantaneous and convolutive mixtures with different number of sources.

Results show that we can build methods based on this framework that outperform the well-known DUET method by 10 dB of SDR in some configurations and have better performance than the multichannel NMF method. Moreover, experiments show that results on convolutive mixture are significantly worse than on instantaneous mixtures. This is not surprising since the mixing model is based on the narrowband approximation which gets poor in reverberant conditions [34]. However, so as to better account for reverberation, the proposed approach could be extended to model the spatial source images, in a similar way as [35].

The proposed model and inference algorithm could also be used for other tasks than BSS such as indexing or audio transcription. Other approximate inference methods could be used to estimate the GSMM in a tractable way. Especially, it would be interesting to consider approximation techniques like variational Bayes [36, 20] to jointly estimate the GSMM and the mixing parameters.

## Acknowledgments

We thank Emmanuel Vincent for kindly sharing his code for the LGM method [10, 37].

## References

- [1] A. Belouchrani and M. Amin, “Blind source separation based on time-frequency signal representations,” *IEEE Transactions on Signal Processing*, vol. 46, no. 11, pp. 2888–2897, 1998.
- [2] P. Bofill and M. Zibulevsky, “Underdetermined blind source separation using sparse representations,” *Signal processing*, vol. 81, no. 11, pp. 2353–2362, 2001.
- [3] O. Yilmaz and S. T. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
- [4] P. O’Grady, B. Pearlmutter, and S. Rickard, “Survey of sparse and non-sparse methods in source separation,” *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 18–33, 2005.
- [5] L. Parra and C. Spence, “Convolutive blind source separation of non-stationary sources,” *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

- [6] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev, "Blind source separation by sparse decomposition in a signal dictionary," in *Independent Component Analysis : Principles and Practice*. Cambridge Press, 2001, pp. 181–208.
- [7] E. Vincent, "Complex nonconvex  $l_p$  norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2007, pp. 430–437.
- [8] C. Févotte and S. Godsill, "A bayesian approach for blind separation of sparse sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2174–2188, Nov. 2006.
- [9] M. E. Davies and N. Mitianoudis, "Simple mixture model for sparse over-complete ICA," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 151, no. 1, pp. 35–43, 2004.
- [10] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined audio source separation via gaussian local modeling," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, 2009.
- [11] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," *Proc. ICA*, pp. 957–961, 2003.
- [12] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1564–1578, July 2007.
- [13] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 191 – 199, jan. 2006.
- [14] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.
- [15] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 5, pp. 3617–3620 vol.5, Apr 1997.
- [16] J. Berger, *Statistical decision theory and Bayesian analysis*. Springer, 1985.
- [17] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 412– 421, 2005.
- [18] Z. Ghahramani and M. Jordan, "Supervised learning from incomplete data via the EM approach," in *Advances in Neural Information Processing Systems (NIPS)*, 1994.

- [19] T. Virtanen, A. Mesaros, and M. Ryyanen, “Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music,” in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition SAPA*, 2008.
- [20] H. Attias, “Independent factor analysis,” *Neural Comput.*, vol. 11, no. 4, pp. 803–851, 1999.
- [21] ———, “New EM algorithms for source separation and deconvolution,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’03)*, 2003, pp. 297–300.
- [22] A. Ozerov, E. Vincent, and F. Bimbot, “A general modular framework for audio source separation,” in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA’10)*, Saint-Malo, France, Sep. 27-30 2010, pp. 33–40.
- [23] Z. Ghahramani and M. I. Jordan, “Factorial hidden Markov models,” *Machine Learning*, vol. 29, pp. 245–273, 1997.
- [24] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, “Blind spectral-GMM estimation for underdetermined instantaneous audio source separation,” in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA’09)*, 2009, pp. 751–758.
- [25] C. Bishop *et al.*, *Pattern recognition and machine learning*. Springer New York:, 2006.
- [26] B. Picinbono, “On circularity,” *Signal Processing, IEEE Transactions on*, vol. 42, no. 12, pp. 3473–3482, 2002.
- [27] A. Ozerov, “Adaptation de modèles statistiques pour la séparation de sources mono-capteur. Application à la séparation voix/musique dans les chansons,” Ph.D. dissertation, 2006.
- [28] E. Vincent, S. Araki, and P. Bofill, “The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation,” in *Proc. of International Conference on Independent Component Analysis and Signal Separation*. Springer, 2009.
- [29] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [30] V. Pulkki and M. Karjalainen, “Localization of amplitude-panned virtual sources I: stereophonic panning,” *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 739–752, 2001.
- [31] D. Campbell, K. Palomaki, and G. Brown, “Roomsim, a matlab simulation of shoebox room acoustics for use in teaching and research,” *Computing and Information Systems*, vol. 9, no. 3, pp. 48–51, 2005.
- [32] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, “First stereo audio source separation evaluation campaign: Data, algorithms and results,” *Lecture Notes in Computer Science*, vol. 4666, p. 552, 2007.

- [33] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures. with application to blind audio source separation,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3137–3140.
- [34] N. Duong, E. Vincent, and R. Gribonval, “Under-determined convolutive blind source separation using spatial covariance models,” in *Acoustics, Speech and Signal Processing, 2010. ICASSP 2010. IEEE International Conference on*, 2010.
- [35] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, “Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation,” in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*, May 2010, pp. 1–4.
- [36] T. Jaakkola and M. Jordan, “Bayesian parameter estimation via variational methods,” *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, 2000.
- [37] N. Duong, E. Vincent, and R. Gribonval, “Spatial covariance models for under-determined reverberant audio source separation,” in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*. IEEE, 2009, pp. 129–132.

## Annexe

	drums				mono				poly			
	3 src	4 src	5 src	6 src	3 src	4 src	5 src	6 src	3 src	4 src	5 src	6 src
DUET	7.65	4.40	1.47	-0.90	16.32	12.85	10.37	5.70	10.01	6.41	4.00	1.49
LGM	10.69	4.86	1.61	-0.96	18.82	13.31	10.37	5.71	12.76	7.21	4.40	1.61
LGM-GMM	12.95	6.46	2.81	-0.18	23.01	16.15	12.68	7.02	15.74	9.24	6.07	2.33
LGM-GSMM	12.46	6.42	2.77	-0.30	22.75	16.38	12.72	6.87	16.13	9.37	6.00	2.58
LGM-GMM-GMM	13.52	6.96	3.01	0.05	24.59	16.74	13.04	7.16	16.23	9.31	6.23	2.37
LGM-GSMM-GMM	14.28	6.82	2.86	-0.12	25.66	17.79	13.24	7.19	16.98	9.50	6.30	2.74
LGM-GSMM-GSMM	14.51	6.60	2.86	-0.34	25.61	17.79	13.26	7.15	<b>17.48</b>	9.89	6.38	2.72
LGM-NMF	12.24	6.37	1.94	-0.58	22.77	15.73	12.58	6.46	15.08	9.61	6.06	2.61
LGM-NMF-NMF	12.78	6.68	2.19	-0.31	24.27	16.52	13.44	6.89	16.09	10.10	6.47	2.79
LGM-NMF-GMM	14.48	7.96	2.81	0.18	24.68	17.14	13.40	7.36	16.58	9.68	6.32	2.31
LGM-MNMF	13.86	7.43	<b>4.09</b>	<b>1.50</b>	25.69	19.58	16.12	9.58	16.87	10.77	7.48	3.59
LGM-MNMF-GMM	14.81	8.98	4.01	1.20	26.07	20.70	16.37	9.67	17.08	10.82	7.77	3.40
LGM-MNMF-GSMM	<b>15.28</b>	<b>9.37</b>	3.90	0.96	26.08	<b>21.88</b>	<b>16.82</b>	<b>10.81</b>	17.31	<b>11.58</b>	<b>7.97</b>	3.43
LGM-MNMF-NMF	14.03	8.30	3.50	0.92	<b>26.12</b>	19.79	15.98	9.39	17.37	11.27	7.91	<b>3.84</b>

Table 1: Source separation results in terms of SDR (dB) on instantaneous mixtures.

	drums				mono				poly			
	3 src	4 src	5 src	6 src	3 src	4 src	5 src	6 src	3 src	4 src	5 src	6 src
DUET	6.16	4.03	2.03	0.99	13.08	10.13	8.69	5.32	9.60	7.10	5.43	2.88
LGM	6.24	4.26	2.37	0.88	12.93	10.88	9.60	6.01	11.06	8.49	6.66	3.65
LGM-GMM	6.73	4.64	2.61	-0.25	13.82	11.87	10.12	6.07	12.19	9.26	6.88	3.07
LGM-GSMM	6.90	4.67	2.61	0.01	13.82	11.90	10.22	6.24	12.28	9.51	6.91	3.37
LGM-GMM-GMM	5.97	3.95	2.57	0.51	8.53	7.41	6.45	3.80	8.90	7.32	5.69	3.02
LGM-GSMM-GSMM	7.02	4.81	2.86	0.35	13.93	12.23	10.46	6.32	12.60	9.69	7.12	3.52
DUET-MNMF	7.08	5.25	3.20	2.11	13.51	11.02	10.83	7.22	11.71	9.78	7.72	4.44
DUET-MNMF-GMM	7.92	6.02	3.83	1.99	13.95	11.53	11.01	7.44	11.90	10.08	7.98	4.36
DUET-MNMF-GSMM	7.75	6.01	3.79	-2.38	13.94	11.56	11.05	7.38	12.08	10.33	7.99	4.43
DUET-MNMF-GSMM-GSMM	7.82	6.31	4.18	2.38	13.96	11.85	11.17	7.54	12.11	10.42	8.18	4.56
LGM-MNMF	7.98	6.46	4.35	2.23	14.62	13.17	<b>12.32</b>	7.98	12.72	10.70	8.78	5.04
LGM-MNMF-GMM	8.00	6.73	4.82	2.51	14.63	13.19	12.10	8.12	12.95	10.67	8.78	4.94
LGM-MNMF-GSMM	<b>8.11</b>	6.82	4.90	2.55	<b>14.73</b>	13.40	12.25	8.16	<b>13.14</b>	10.80	8.92	4.99
LGM-MNMF-GSMM-GSMM	8.06	<b>6.83</b>	<b>5.07</b>	<b>2.79</b>	14.70	<b>13.48</b>	12.27	<b>8.22</b>	13.10	<b>10.85</b>	<b>9.10</b>	<b>5.13</b>

Table 2: Source separation results in terms of SDR (dB) on convolutive mixtures.



---

Centre de recherche INRIA Rennes – Bretagne Atlantique  
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex  
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier  
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq  
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique  
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex  
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex  
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex  
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

---

Éditeur  
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)  
<http://www.inria.fr>  
ISSN 0249-6399