A tractable framework for estimating and combining spectral source models for audio source separation

Simon Arberet, Alexey Ozerov, Frédéric Bimbot, Rémi Gribonval

A Tractable Framework for Estimating and Combining Spectral Source Models for Audio Source Separation [1]

Simon Arberet[a], Alexey Ozerov[b], Frédéric Bimbot[c], Rémi Gribonval[b]

[a] *Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.*
e-mail: simon.arberet@epfl.ch
[b] *INRIA, Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes cedex, France.*
e-mails: {alexey.ozerov,remi.gribonval}@inria.fr
[c] *IRISA, CNRS - UMR 6074, Campus de Beaulieu, 35042 Rennes cedex, France.*
e-mail: frederic.bimbot@irisa.fr

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 Scaled 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 problems 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 different probabilistic models 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.

Keywords: Blind source separation, multichannel audio, Gaussian mixture model, expectation-maximization algorithm, convolutive mixture.

[1] 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.

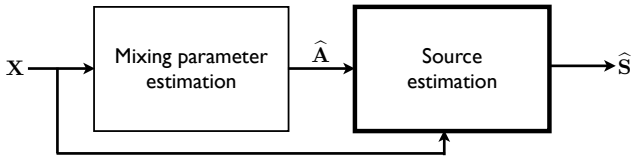


Figure 1: Block diagram of the two step approach. In this paper, we focus on the second block i.e. the source estimation.

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:

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

where $s_n(\tau)$ and $x_m(\tau)$ denote sampled time signals of respectively the n-th source and the m-th mixture (τ 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 $x_m(\tau)$ $(1 \leq m \leq M)$.

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:

- at the first step the mixing parameters are estimated [3, 5, 6, 7], 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 (by the so-called narrowband approximation [8]) as follows [9] [2]:

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

where $\mathbf{S}(t,f) = [S_1(t,f),\ldots,S_N(t,f)]^T$ and $\mathbf{X}(t,f) = [X_1(t,f),\ldots,X_M(t,f)]^T$ ($t = 1,\ldots,T$ and $f = 1,\ldots,F$ being time and frequency indices) denote respectively column vectors of source and mixture STFTs; and $\mathbf{A}(f) = [A_{mn}(f)]_{m,n=1}^{M,N}$ is an $M \times N$ complex-valued mixing matrix with elements $A_{mn}(f)$ being the discrete Fourier transforms of filters $\mathbf{a}_n(l) = [a_{1n}(l),\ldots,a_{Mn}(l)]^T$.

Sparse prior distributions that are mostly used for audio sources include Laplacian [10, 2], generalized Gaussian [11], Student-t [12], and mixtures of two Gaussians [13]. One of the most popular classical sparsity-based approaches is the DUET method [3]. It assumes that at each TF point, there is approximately one active source. The source that is supposed to be the active one, is the one that has its direction closest to the mixture observation $\mathbf{X}(t,f)$. The estimated value of this source is the projection of the mixture vector $\mathbf{X}(t,f)$ on the corresponding source direction. The estimated value of the other sources is zero. The underlying principles of other sparsity-based two-step approaches [1, 2, 4] are quite similar to that of DUET, and all these methods [1, 2, 3, 4] suffer from the following common limitations:

1. in each TF point at most M of N (recall that $M < N$) source coefficients are reconstructed with nonzero values, which leads to errors in the TF points where there are more than M nonzero sources,

2. each TF coefficient is assumed to be independent of the others, and, as a consequence, the redundancy and structure of audio sources are not taken into account.

The first issue and partially the second one have been addressed by the Local Gaussian Model (LGM) [14], where source TF coefficients are locally modeled by Gaussian distributions with free variances. This method allows (in the instantaneous case) reconstructing up to $M(M+1)/2$ source coefficients with nonzero values. However, this method exploits only a neighborhood of each TF point, in order to estimate the parameters of the corresponding Gaussian distribution.

[2] 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.

Number of sources (N)	3	4	5	6
Number of components in the observation ($\mathcal{K}^N$)	512	4 096	32 768	262 144
Time per EM iteration	1 min.	6 hours	N.A.	N.A.
Virtual memory size	3 GB	9 GB	N.A.	N.A.

Table 1: Complexity of the global EM algorithm for a GMM model with $\mathcal{K} = 8$ components per source. To compute the time per iteration, we used MATLAB on an Intel Core i7 processor running at 2.66 GHz with 4 GB of RAM. The computation for 4 sources were so long that we did not even try to run the algorithm for 5 and 6 sources.

A more globally structured approach consists in assuming a structured spectral model of each source, such as Spectral Gaussian Mixture Model (Spectral-GMM) [15, 16], the Spectral Gaussian Scaled Mixture Model (Spectral-GSMM) [17] or the Spectral Nonnegative Matrix Factorization (Spectral-NMF) model [18]. The Spectral-GMM and Spectral-GSMM have been successfully used to separate sources in the single channel case ($M = 1$) [15, 16, 17], 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 [19], 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. If there are N sources and $\mathcal{K}$ components per source, the number of components in the observation is $\mathcal{K}^N$. For a typical value of $\mathcal{K} = 8$ the resulting computational complexity and memory requirements for 3 to 6 sources increases very fast as depicted in Table 1. It shows that there is a clear limit in the number of sources that can be estimated using the global EM algorithm for a state based model.

- *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 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, which we hope will better match the sources and thus lead to a better estimator of the sources. 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. Contribution 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 [20], and then learning the source models from the source estimates $\tilde{S}_n$ while taking into account the noise.

This framework allows computational complexity to be reduced 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. Contribution 2: Multi-layer "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 different probabilistic models (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. In other words, given that different models assume different probability distributions of source coefficients, it becomes impossible to use them in a joint manner, while this kind of multi-layer fusion allows using different models

in a sequential way, thus making some profit from all of them. 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 (e.g., LGM), and then, learn a model (on the next layer) which better represents the source characteristics (e.g., Spectral-GSMM). In fact, we show in the experimental part that among different tested settings the best results were obtained using a multi-layer combination of different models.

1.1.3. Generality and summary of contributions

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* [21], where a Gaussian model is used to model the error of clean speech features estimation from noisy speech. In summary this framework offers:

- *contribution 1:* a computationally tractable approximate inference in factorial state-based models; and
- *contribution 2:* the ability to combine probabilistic models 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 sensitivity to initialization.

1.2. Related works

The proposed framework is related to that of learning GMMs from incomplete data by Ghahramani and Jordan [22], and extends it in the sense that [22] 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 [23], 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 [22]. As compared to [23], 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 [24, 25] 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 [26] [3], 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 [22] and variational inference in factorial state-based models in the single [27] and multi-channel [24] 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 of Fig. 1, "source estimation"), 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 [28] 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), and we provide a more consistent experimental evaluation.

[3] Note that in [26] there is no combination of models in a multi-layer fashion.

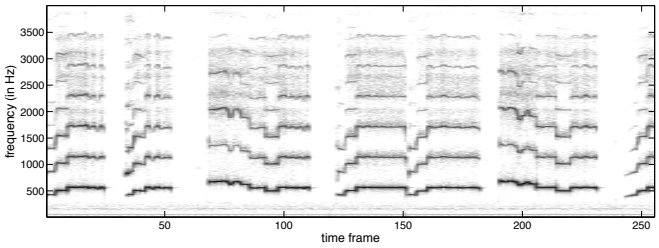
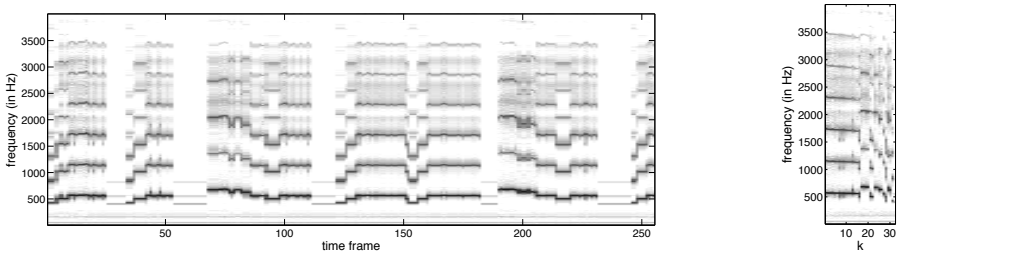


Figure 2: Spectrogram of a flute performance. The sampling frequency is 8 kHz and the window size is 1024 samples. The amplitude of each TF point is represented as shades of gray varying from black for the strongest amplitude to white for the weakest.



(a) Decoded spectrogram $V_{k(t)}, 1 \leq t \leq T$. (b) Spectral patterns $V_k, 1 \leq k \leq \mathcal{K} = 32$

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

2. Source models

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

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 structured spectral models which we describe hereafter.

LGM [14], Spectral-GMM [15, 16], Spectral-GSMM [15, 17] and Spectral-NMF [18] are all zero-mean Gaussian [4] source models, but with different assumptions on the structure of the

[4] 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.

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 imposes some structure on 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.

2.1. Local Gaussian Model (LGM)

In the LGM model [14], the source TF coefficients $S(t,f)$ are assumed to be realizations of independent zero-mean complex-valued Gaussian variables with variances $\sigma^2(t,f)$:

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

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:

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

where "$(.)^H$" denotes conjugate-transpose, μ is the p-dimensional complex-valued mean vector and Σ is the $p \times p$ complex-valued covariance matrix.

In this simple local model, the variance represents the local Power Spectral Density (PSD) of the source and the phase is assumed uniformly random (in $[0\ 2\pi]$).

The variance $\sigma^2(t,f)$ can be estimated in the neighborhood of the corresponding TF point using some symmetrical overlapping two-dimensional window [14].

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 $S(t,f)$, $f = 1,\ldots,F$ 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:

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

where $\bar{0}$ is a vector of all zeros, π_k (satisfying $\sum_{k=1}^{\mathcal{K}}\pi_k = 1$) and Σ_k denote respectively the weight and the diagonal covariance matrix of the k-th GMM state. Introducing nonnegative

column vectors of variances $V_k \triangleq [\sigma_k^2(f)]_f$, the k-th state covariance matrix can be expressed as $\Sigma_k = \mathrm{diag}\,(V_k)$. 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 $V_{k(t)} \in \mathbb{V} = \{V_1, \ldots, V_{\mathcal{K}}\}$ 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 $\sigma^2_{k(t)}(f)$:

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

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 θ^{gmm} the *complete* model of the Spectral-GMM λ^{gmm}.

As opposed to the LGM model, where there are as many free parameters (variances) as TF points (this is called a *non-parametric* model in statistical settings), the Spectral-GMM is defined by $\mathcal{K} \times F$ free parameters (variances) only (this is called a *semiparametric* model in statistical settings). Figure 3(b) depicts the spectral patterns of a 32-state 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 to the original spectrogram of Figure 2. As opposed to the LGM, where the variance of each TF point is estimated independently, the Spectral-GMM exploits the global structure of the signal. As a consequence, fewer free parameters are to be estimated and thus their estimation is likely to be statistically more consistent.

However, the GMM does not explicitly model the 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 [17].

2.3. Spectral Gaussian Scaled Mixture Models (GSMM)

The GSMM model [17] 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:

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

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 $V_{k(t)} \in \mathbb{V} = \{V_1, \ldots, V_{\mathcal{K}}\}$ 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 $g_k(t)\sigma^2_{k(t)}(f)$:

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

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

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

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 as follows:

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

where $V = \left[v_k(f)\right]_{f,k} = [V_1, \ldots, V_{\mathcal{K}}]$ and $H = [h_k(t)]_{k,t}$ are two non-negative matrices [18]. 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 becomes intractable as the number of sources increases. We then present our approach in section 3.2.

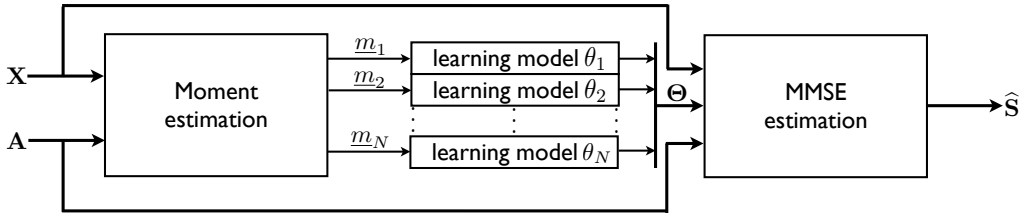


Figure 4: Block diagram of the proposed BSS framework. $\mathbf{X}$ is the mixture, $\mathbf{A}$ is the mixing matrix, $\underline{m}_n = \left\{m_{1,n}, m'_{2k,n}\right\}$ are the posterior moments of source n defined in (12) and (13), $\mathbf{\Theta} = [\theta_n]_{n=1}^N$ is the concatenation of the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

3.1. Learning from the mixture

Given the Spectral-GSMM [5] source models $\mathbf{\Lambda}^{\text{gsmm}} = [\lambda_n^{gsmm}]_{n=1}^N$ 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 [24]. 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* [24, 29]) that $\mathbf{k} \triangleq [k_n]_{n=1}^N$ is the active state at frame t:

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

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 (i.e. LGM described in section 2.1), based on a source model λ^{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):

[5] Spectral-GMM is a special case of Spectral-GSMM with all the gains fixed to one.

$$m_1 = \mathbb{E}\left\{S|\mathbf{X}, \lambda^{init}\right\} \tag{12}$$

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

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 zero [30]. 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:

$$\tilde{S} = S + E, \tag{14}$$

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 λ^{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 σ_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 λ (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 [15, 16] with observed mixture $\widetilde{S}$ and two latent sources: *target source* S and *noise* E. The model is learned by optimizing the ML criterion:

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

where λ_e is the noise model whose 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:

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

As mentioned in sections 2.2 and 2.3, given the state-decoded sequence $\{k(t)\}_t$ and the source model parameters λ, the source can be assumed to be distributed according to a simpler model (a Gaussian distribution if λ 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 θ as:

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

As a consequence, θ denotes a model that can be considered as no-state based, and thus its MMSE source estimates is not combinatorial. In the case of the GMM, the MMSE of the complete model θ^{gmm} is sometime called the "hard" estimator while the MMSE of model λ^{gmm} is called the "soft" estimator [31] .

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 $\mathbf{\Theta} = [\theta_n]_{n=1}^N$ with the MMSE estimator detailed in the next section which will be detailed in section 4.

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 λ 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 $\mathbf{\Theta}$.

Let assume that, given the model $\mathbf{\Theta}$ 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

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

If $\mathbf{\Theta} \triangleq \mathbf{\Lambda} = [\lambda_n]_{n=1}^N$, 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 $\mathbf{\Theta} \triangleq \{\mathbf{\Lambda}, \mathbf{\Gamma}\}$, i.e. if the state decoding sequence $\mathbf{\Gamma} = [\mathbf{k}(t)]_{t=1}^T$ is included into the set of model parameters $\mathbf{\Theta}$ 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 $\delta(.)$ that minimizes the posterior risk $r(t,f) = \mathbb{E}\{L(\mathbf{S}(t,f), \delta(\mathbf{X}(t,f)))|\mathbf{X}(t,f),\mathbf{\Theta}\}$, with $L(.,.)$ 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 $\delta_{\mathbf{\Theta}}(\mathbf{X}(t,f)) = \mathbb{E}\{\mathbf{S}(t,f)|\mathbf{X}(t,f),\mathbf{\Theta}\}$.

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 source 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:

$$\mathbf{m}_1(t,f) \triangleq \delta_{\mathbf{\Theta}}(\mathbf{X}(t,f)) \tag{18}$$

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

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

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

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

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

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

The resulting posterior covariance matrix is:

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

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* [6] 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 $\mathbf{\Sigma}(t,f) = \mathbb{E}[\mathbf{S}(t,f)\mathbf{S}(t,f)^H|\mathbf{\Theta}]$ which is diagonal since the sources are assumed independent. The diagonal entries of $\mathbf{\Sigma}(t,f)$ 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 ($\lambda \triangleq \lambda^{gsmm}$) and in section 5.2 we derive an EM algorithm in the case where the source model is a Spectral-NMF ($\lambda \triangleq \lambda^{nmf}$).

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 $\Sigma_e(t) = \mathrm{diag}([\sigma_e^2(t,f)]_f)$. Then we only have to match $m'_2(t,f)$ to the noise variances $\sigma_e^2(t,f) \triangleq m'_2(t,f)$.

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 [16]. The Spectral-GMM model being a special case of the Spectral-GSMM with gains equal to one, the EM algorithm for Spectral-GMM is the one described in [16], that is the same as Algorithm 1 without step 4 and with $g_k^{(l)}(t) = 1, \forall k,t,l$.

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

It can easily be checked that algorithm 1 is more general than the EM algorithm for learning GMMs from incomplete data [22] in the sense that if a subset of noise variances tends to zero

[6] see its definition in section 2.2

(observed TF points), and the rest of the variances to $+\infty$ (missing TF points), Algorithm 1 and the EM algorithm from [22] 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

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

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

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

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

3. Re-estimate Gaussian weights

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

4. Re-estimate Gaussian gains

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

5. Re-estimate covariance matrices

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

5.1.1. Decoding step

Once the source model $\mathbf{\Lambda}^{\text{gsmm}} = [\lambda_n^{gsmm}]_{n=1}^N$ have been 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 responsibilties 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 $\mathbf{\Lambda}^{\text{gsmm}}$, i.e., $\delta_{\mathbf{\Lambda}^{\text{gsmm}}}(\mathbf{X}(t,f)) = \sum_{\mathbf{k}} \gamma_{\mathbf{k}}(t)\mathbf{W}_{\mathbf{k}}(t,f)\mathbf{X}(t,f)$ which implies again computing a sum over the $\mathcal{K}^N$ states, we consider the complete model $\mathbf{\Theta}^{gsmm} \triangleq \{\mathbf{\Lambda}^{\text{gsmm}}, \mathbf{\Gamma}\}$, with $\mathbf{\Gamma} = [\mathbf{k}^*(t)]_{t=1}^T$ and $\mathbf{k}^*(t) = [k_n^*(t)]_{n=1}^N$. That is, we keep only the most likely state given by:

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

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

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

5.2. Learning Spectral-NMF from noisy observations

Algorithm 2 summarizes an EM algorithm for the optimization of criterion (15) when λ 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 [18].

6. Experimental study

We evaluate the framework described in section 3 using LGM as an initial BSS method ($\lambda = \lambda^{init}$) 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* [7], *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.

[7] 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 [32]).

Algorithm 2 EM Algorithm for source Spectral-NMF estimation in the ML sense (index (l) in power denotes the parameters estimated at the l^{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)

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

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 [18].

The goals of the experiments provided below are to investigate the potential of the proposed framework in the single and multi-layer settings (Experiments 2 and 3) and to provide experimental support to two contributions presented in the introduction, notably (1) tractable approximate inference and (2) multi-layer model fusion (Experiments 1 and 4). More precisely, we propose four experiments aiming at:

- *Experiment 1:* Comparison with the global EM approach. We compare the computational complexity and the performance of our decoupled EM approach with the global EM approach discussed in the introduction.
- *Experiment 2:* 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 3:* 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;

- *Experiment 4:* evaluating the performance of the framework as compared with the state of the art. Thus we want to compare our framework with other structured source model based approaches like LGM and Multichannel-NMF. This experiment is first performed for the case of instantaneous mixtures and then it is repeated in Section 6.3 for convolutive mixtures.

6.1. Experimental setting

6.1.1. Experimental material

We evaluate the methods over stereo mixtures ($M = 2$) [8] 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 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 [33]. The results in this section are given in SDR gain w.r.t. the popular classical DUET method [3]. However the absolute source separation performance in SDR of all tested methods are shown in the tables in the appendix.

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

[8] The stereo case ($M = 2$) is the most common in audio, and also it is the most challenging among the multichannel settings. Thus, while the framework itself is not limited to the stereo case, we do not consider here the cases of more than two channels ($M > 2$).

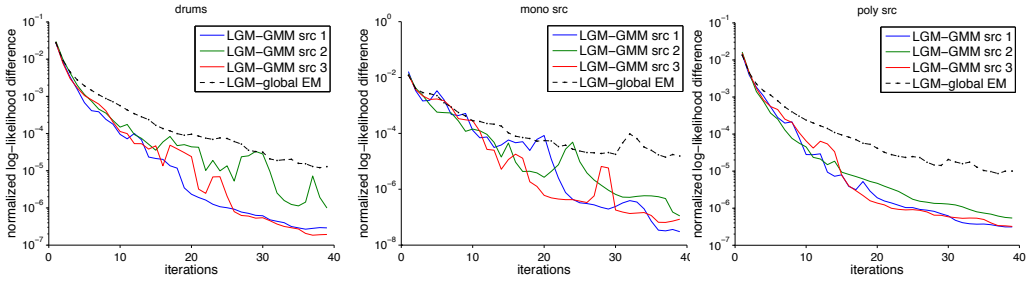
10 by 1 rectangular window in the convolutive case. The size of these neighborhoods yielded the best results on our preliminary experiments on respectively instantaneous and convolutive mixtures. The Spectral-GSMMs (including the Spectral-GMMs) and Spectral-NMF models are learned with 40 iterations of, 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 [34], given an angle of $50 - 5N$ degrees between successive sources.

6.2.1. Results

Experiment 1: Comparison with the global EM approach. As mentioned in the introduction, one of our contributions is to find an alternative to the global EM approach because of its computational complexity when the model is a state based model such as a GMM or GSMM. To show the benefit of our decoupled approach compared to the global EM approach, we compare the computational complexity and the performance of both of these approaches when the source model is a GMM and the mixture is composed of 3 sources. We don't make any comparison when the number of sources is higher because, as illustrated in Table 1, the computational complexity of the global EM approach in these cases is too high. Results in Table 2 shows that in terms of performance, the results of the global EM approach and the decoupled EM approach are similar, but in terms of computational complexity, the decoupled EM approach is far more efficient: The time to compute each iteration is 600 times smaller with our decoupled approach. We can also wonder what is the speed of convergence with respect to the number of iterations. We have to be careful when doing this comparison because the two approaches (global one and the decoupled one) are based on two different likelihood expressions. Thus, to be able to compare these two approaches, we computed the normalized difference of the log-likelihood $nlld(i)$ between the next iteration $i + 1$ and the current iteration i. i.e. $nlld(i) = (ll(i + 1) - ll(i)) / ll(I)$, where $ll(i)$ is the log-likelihood at iteration i and I is the last iteration computed. Here $I = 40$. When the normalized differences get close to zero means that the algorithm get close to convergence.



(a) drums sources (b) monophonic sources (c) polyphonic sources

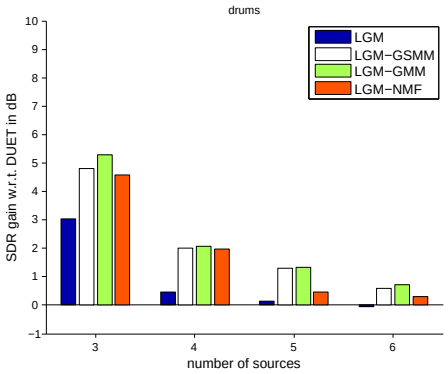
Figure 5: Average (over the 20 examples of each database) normalized log-likelihood difference $nlld(i)$ between the next iteration $i+1$ and the current one i (i.e. $nlld(i) = (ll(i+1) - ll(i))\,/ll(I)$, where $ll(i)$ is the log-likelihood at iteration i and I is the last iteration computed. Here $I = 40$.), of the global EM (black dotted line) and the proposed approach (the three colored lines) which runs on each source separately.

Figure 5 shows that the *nlld* plot of the global EM algorithm is above the one of the decoupled approach, showing that the convergence is slower with respect to the number of iterations. This experiment clearly supports our first contribution (tractable approximate inference), since the proposed decoupled EM leads to similar separation performance as the global EM, while being much faster.

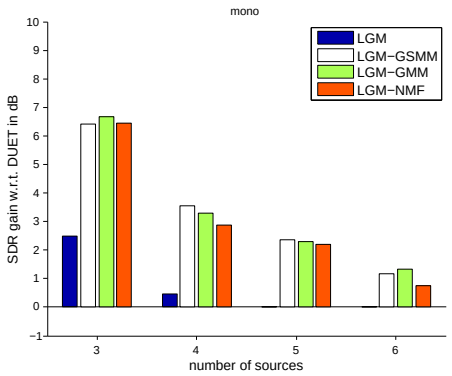
	SDR			time per iteration
	Drums	Monophonic	Polyphonic	
LGM - GMM global EM	12.75 dB	23.29 dB	14.98 dB	1 minute
LGM - GMM (decoupled EM)	12.95 dB	23.01 dB	15.74 dB	0.1 second

Table 2: Performance comparison of the global EM algorithm with our decoupled approach when there are 3 sources. The source model is a GMM, the number of components per source is $\mathcal{K} = 8$, the number of iterations is 40. To compute the time per iteration, we used MATLAB on an Intel Core i7 processor running at 2.66 GHz with 4 GB of RAM.

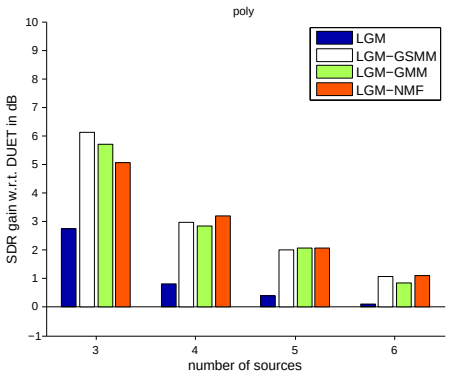
Experiment 2. Figure 6 compares the average SDR (improvement w.r.t. DUET) achieved by LGM and the proposed LGM-{GMM,GSMM,NMF} methods. The results show that running one layer of the proposed framework with any of the considered models {GMM,GSMM,NMF} improves the SDR of LGM by severals dBs, especially on the monophonic dataset when there are few sources. For instance, there is approximately 3 dBs of improvement w.r.t. LGM and 6



(a) drums sources



(b) monophonic sources

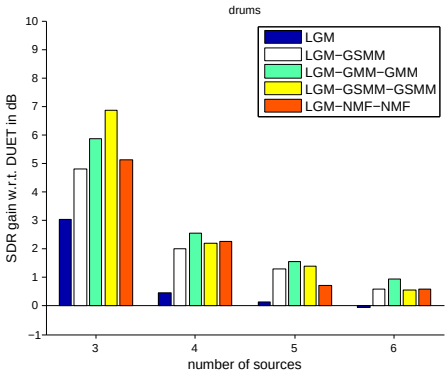


(c) polyphonic sources

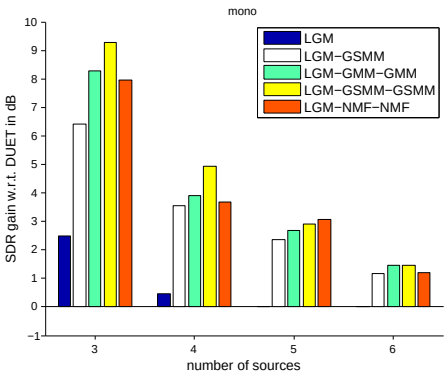
Figure 6: Experiment 1: Source separation gain w.r.t. DUET of LGM and some methods based on our proposed one-layer framework: LGM-GMM, LGM-GSMM and LGM-NMF on instantaneous mixtures with respect to the number of sources.

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.

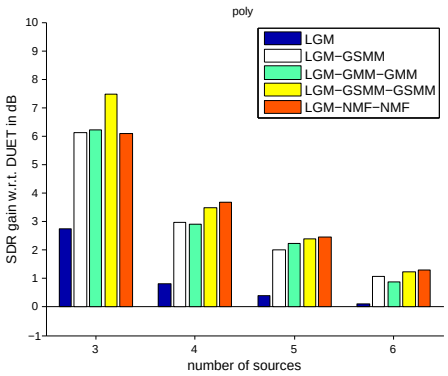
Experiment 3. Figure 7 compares the average SDR (improvement w.r.t. DUET) achieved by LGM and single-layer (e.g. LGM-GSMM) and two-layer (e.g. LGM-GSMM-GSMM) methods of our framework. For the sake of legibility, we only plot the results of one of the single layer methods that showed the best result within 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 severals 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



(a) drums sources



(b) monophonic sources

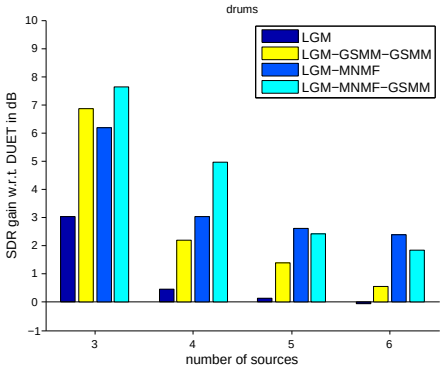


(c) polyphonic sources

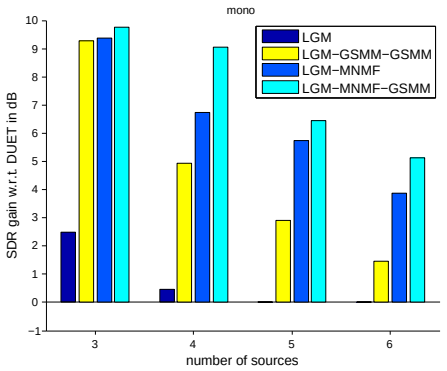
Figure 7: Experiment 2: Source separation gain w.r.t. DUET of LGM and some methods based on our two-layer framework: LGM-GMM-GMM, LGM-GSMM-GSMM and LGM-NMF-NMF on instantaneous mixtures with respect to the number of sources.

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.

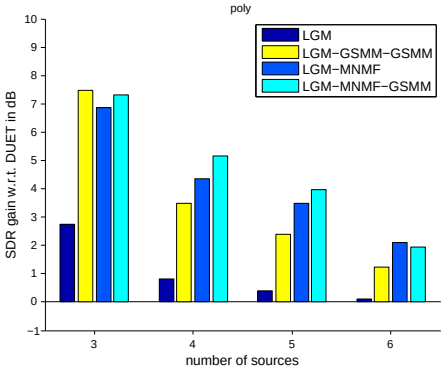
Experiment 4. Figure 8 compares the average SDR (improvement w.r.t. DUET) achieved by classical methods like LGM and Multichannel-NMF with LGM initialization (LGM-MNMF) with LGM-GSMM-GSMM and LGM-MNMF-GSMM (i.e. GSMM learned with our framework on the top of LGM-MNMF). We can notice that LGM-MNMF performs 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



(a) drums sources



(b) monophonic sources



(c) polyphonic sources

Figure 8: 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.

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. Finally, these results support the benefit of the multi-layer model fusion, since the best performance was obtained using different models (LGM-MNMF-GSMM), thus approving our second contribution.

6.3. Convolutive mixtures

Now we compare the same methods as in the previous section on synthetic convolutive mixtures, i.e., static sources filtered by synthetic room impulse responses simulating a pair of omnidirectional microphones via the Roomsim toolbox [35]. The room dimensions are: 4.45 x 3.55 x

2.5 m. Depending on the mixture, the source directions of arrival vary between -60 degrees and +60 degrees with a minimal spacing of 15 degrees and the distances between the sources and the center of the microphone pair is 1 m. The distance between the two microphones is 1 m and all the sources and microphones are at a height of 1.4 m. The reverberation time (RT) is set to 130 ms while the sampling frequency is 16 kHz.

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 [36], that is:

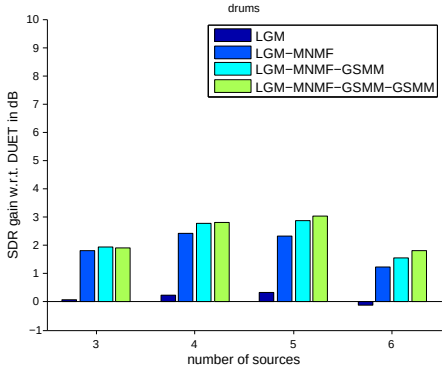
$$s_{mn}^{img}(\tau) = a_{mn} * s_n(\tau)$$

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

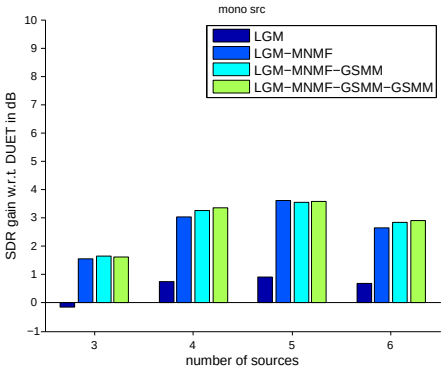
Note that the DUET [3] and the Multichannel NMF [18] also address the mixing matrices estimation. Here, as we evaluate only the source estimation task, we consider that $\mathbf{A}(f)$ is the same for all the tested methods and is given by the above mentioned PCA procedure.

6.3.2. Results

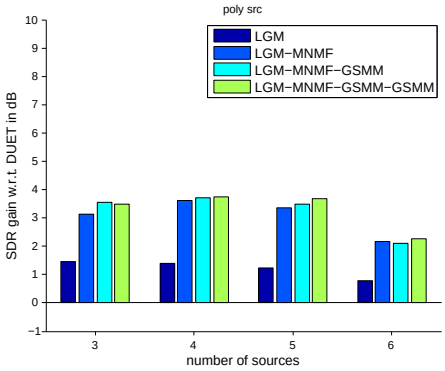
The results on convolutive mixtures are depicted in Figure 9 and Tables 4. Unsurprisingly, the absolute performance of all the tested methods are worse than in the instantaneous case. However the methods based on a structured spectral models (LGM-MNMF,LGM-MNMF-GSMM, LGM-MNMF-GSMM-GSMM) improve the SDR by several dBs. Adding one or two layers of GSMM to LGM-MNMF slightly increases the performance (up to 1 dB depending on the configuration). It is also interesting to notice that, as opposed to the instantaneous case, the performance variation with the number of sources is not so pronounced.



(a) drums sources



(b) monophonic sources



(c) polyphonic sources

Figure 9: 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.

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 an 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 (*contribution 1*). Moreover, the proposed framework is multi-layer and can combine different probabilistic models that can be seen as a sort of

probabilistic fusion (*contribution 2*). It is for example possible to learn different source models on the same source while taking into account the error of the source estimate. Thus by refining the source model estimate we also expect to have a better estimation of the sources, which was confirmed by the experiments.

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.

Experimental evaluation supports both contributions claimed in the introduction. First, the proposed approximate inference in state-based models leads to similar separation performance, as the global EM, while it is much faster and tractable. Second, the best performance was obtained using different models combined in the multi-layer fashion, thus supporting the advantage of the proposed multi-layer model fusion. Moreover, 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. While the proposed methods outperforms DUET on convolutive mixture by few dBs of SDR, the overall performance is significantly worse than on instantaneous mixtures. This is not surprising since the mixing model is based on the narrowband approximation which becomes poor in reverberant conditions [37]. However, to better account for reverberation, the proposed approach could be extended to model the spatial source images, in a similar way as [38].

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 [39, 24] to jointly estimate the GSMM and the mixing parameters.

Acknowledgments

We thank Emmanuel Vincent for kindly sharing his code for the LGM method [14, 40].

References

[1] A. Belouchrani, M. Amin, Blind source separation based on time-frequency signal representations, IEEE Transactions on Signal Processing 46 (11) (1998) 2888–2897.

[2] P. Bofill, M. Zibulevsky, Underdetermined blind source separation using sparse representations, Signal processing 81 (11) (2001) 2353–2362.

[3] O. Yılmaz, S. T. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Trans. on Signal Processing 52 (7) (2004) 1830–1847.

[4] P. O'Grady, B. Pearlmutter, S. Rickard, Survey of sparse and non-sparse methods in source separation, International Journal of Imaging Systems and Technology 15 (1) (2005) 18–33.

[5] M. Puigt, Y. Deville, Time–frequency ratio-based blind separation methods for attenuated and time-delayed sources, Mechanical Systems and Signal Processing 19 (6) (2005) 1348–1379.

[6] S. Arberet, R. Gribonval, F. Bimbot, A robust method to count and locate audio sources in a multichannel underdetermined mixture, Signal Processing, IEEE Transactions on 58 (1) (2010) 121 –133.

[7] S. Arberet, P. Sudhakar, R. Gribonval, A wideband doubly-sparse approach for mito sparse filter estimation, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, 2011.

[8] W. Kellermann, H. Buchner, Wideband algorithms versus narrowband algorithms for adaptive filtering in the DFT domain, Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on 2 (2003) 1278–1282 Vol.2.

[9] L. Parra, C. Spence, Convolutive blind source separation of non-stationary sources, IEEE Trans. Speech and Audio Processing 8 (3) (2000) 320–327.

[10] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, 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.

[11] 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.

[12] C. Févotte, S. Godsill, A Bayesian approach for blind separation of sparse sources, Audio, Speech, and Language Processing, IEEE Transactions on 14 (6) (2006) 2174–2188.

[13] M. E. Davies, N. Mitianoudis, Simple mixture model for sparse overcomplete ICA, IEE Proceedings on Vision, Image and Signal Processing 151 (1) (2004) 35–43.

[14] E. Vincent, S. Arberet, R. Gribonval, Underdetermined audio source separation via gaussian local modeling, in: Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA), 2009.

[15] L. Benaroya, F. Bimbot, Wiener based source separation with HMM/GMM using a single sensor, Proc. ICA (2003) 957–961.

[16] A. Ozerov, P. Philippe, F. Bimbot, 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 15 (5) (2007) 1564–1578.

[17] L. Benaroya, F. Bimbot, R. Gribonval, Audio source separation with a single sensor, Audio, Speech, and Language Processing, IEEE Transactions on 14 (1) (2006) 191 – 199.

[18] A. Ozerov, C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, IEEE Transactions on Audio, Speech and Language Processing 18 (3) (2010) 550–563.

[19] E. Moulines, J.-F. Cardoso, 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 5 (1997) 3617–3620 vol.5.

[20] J. Berger, Statistical decision theory and Bayesian analysis, Springer, 1985.

[21] L. Deng, J. Droppo, 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 13 (2005) 412– 421.

[22] Z. Ghahramani, M. Jordan, Supervised learning from incomplete data via the EM approach, in: Advances in Neural Information Processing Systems (NIPS), 1994.

[23] T. Virtanen, A. Mesaros, M. Ryynanen, 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.

[24] H. Attias, Independent factor analysis, Neural Comput. 11 (4) (1999) 803–851.

[25] H. Attias, 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.

[26] A. Ozerov, E. Vincent, 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, 2010, pp. 33–40.

[27] Z. Ghahramani, M. I. Jordan, Factorial hidden Markov models, Machine Learning 29 (1997) 245–273.

[28] S. Arberet, A. Ozerov, R. Gribonval, 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.

[29] C. Bishop, et al., Pattern recognition and machine learning, Springer New York:, 2006.

[30] B. Picinbono, On circularity, Signal Processing, IEEE Transactions on 42 (12) (2002) 3473–3482.

[31] A. Ozerov, Adaptation de modèles statistiques pour la séparation de sources mono-capteur. Application à la séparation voix/musique dans les chansons, Ph.D. thesis (2006).

[32] E. Vincent, S. Araki, 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.

[33] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE Trans. on Audio, Speech and Language Processing 14 (4) (2006) 1462–1469.

[34] V. Pulkki, M. Karjalainen, Localization of amplitude-panned virtual sources I: stereophonic panning, Journal of the Audio Engineering Society 49 (9) (2001) 739–752.

[35] D. Campbell, K. Palomaki, G. Brown, Roomsim, a MATLAB simulation of shoebox room acoustics for use in teaching and research, Computing and Information Systems 9 (3) (2005) 48–51.

[36] E. Vincent, H. Sawada, P. Bofill, S. Makino, J. Rosca, First stereo audio source separation evaluation campaign: Data, algorithms and results, Lecture Notes in Computer Science 4666 (2007) 552.

[37] N. Duong, E. Vincent, 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.

[38] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, 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, 2010, pp. 1 –4.

[39] T. Jaakkola, M. Jordan, Bayesian parameter estimation via variational methods, Statistics and Computing 10 (1) (2000) 25–37.

[40] N. Duong, E. Vincent, 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.

Appendix

	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	**17.48**	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	**4.09**	**1.50**	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	**15.28**	**9.37**	3.90	0.96	26.08	**21.88**	**16.82**	**10.81**	17.31	**11.58**	**7.97**	3.43
LGM-MNMF-NMF	14.03	8.30	3.50	0.92	**26.12**	19.79	15.98	9.39	17.37	11.27	7.91	**3.84**
LGM-GMM global EM	12.75	N.A.	N.A.	N.A.	23.29	N.A.	N.A.	N.A.	14.98	N.A.	N.A.	N.A.

Table 3: Source separation results in terms of SDR (dB) on instantaneous mixtures. The best result of each column is in bold.

	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	**12.32**	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	**8.11**	6.82	4.90	2.55	**14.73**	13.40	12.25	8.16	**13.14**	10.80	8.92	4.99
LGM-MNMF-GSMM-GSMM	8.06	**6.83**	**5.07**	**2.79**	14.70	**13.48**	12.27	**8.22**	13.10	**10.85**	**9.10**	**5.13**

Table 4: Source separation results in terms of SDR (dB) on convolutive mixtures. The best result of each column is in bold.