A General Flexible Framework for the Handling of Prior Information in Audio Source Separation

Alexey Ozerov, Emmanuel Vincent, Frédéric Bimbot

A General Flexible Framework for the Handling of Prior Information in Audio Source Separation

Alexey Ozerov, *Member, IEEE,* Emmanuel Vincent, *Senior Member, IEEE,* and Frédéric Bimbot

***Abstract*—Most of audio source separation methods are developed for a particular scenario characterized by the number of sources and channels and the characteristics of the sources and the mixing process. In this paper we introduce a general audio source separation framework based on a library of structured source models that enable the incorporation of prior knowledge about each source via user-specifiable constraints. While this framework generalizes several existing audio source separation methods, it also allows to imagine and implement new efficient methods that were not yet reported in the literature. We first introduce the framework by describing the model structure and constraints, explaining its generality, and summarizing its algorithmic implementation using a generalized expectation-maximization algorithm. Finally, we illustrate the above-mentioned capabilities of the framework by applying it in several new and existing configurations to different source separation problems. We have released a software tool named** *Flexible Audio Source Separation Toolbox (FASST)* **implementing a baseline version of the framework in Matlab.**

***Index Terms*—Audio source separation, local Gaussian model, nonnegative matrix factorization, expectation-maximization**

I. Introduction

Separating audio sources from multichannel mixtures is still challenging in most situations. The main difficulty is that audio source separation problems are usually mathematically ill-posed and to succeed one needs to incorporate additional knowledge about the mixing process and/or the source signals. Thus, efficient source separation methods are usually developed for a particular source separation problem characterized by a certain *problem dimensionality*, e.g., determined or underdetermined, certain *mixing process characteristics*, e.g., instantaneous or convolutive, and certain *source characteristics*, e.g., speech, singing voice, drums, bass or noise [1]. For example, a source separation problem may be formulated as follows:

> *"Separate bass, drums, melody and the remaining instruments from a stereo professionally produced music recording."*

Given a source separation problem, one typically must introduce as much knowledge about this problem as possible into the corresponding separation method so as to achieve good separation performance. However, there is often no common formulation describing methods applied for different problems, and this makes it difficult to reuse a method for a problem it was not originally conceived for. Thus, given a new source separation problem, the common approach consists in (i) model design, taking into account problem formulation, (ii) algorithm design and (iii) implementation (see Fig. 1, top).

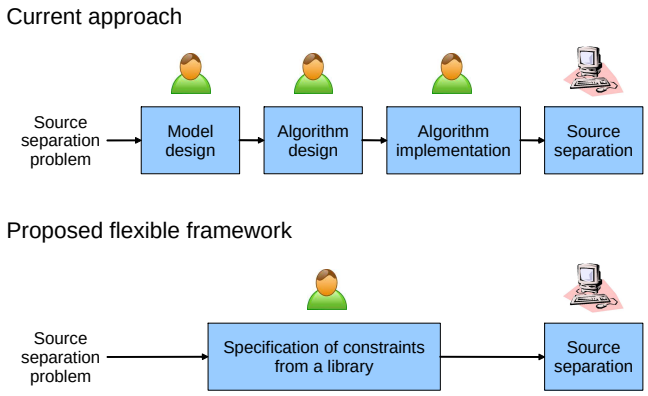

Fig. 1. Current way of addressing a new source separation problem (top) and the way of addressing it using the proposed flexible framework (bottom).

The motivation of this work is to improve over this time-consuming process by designing a general audio source separation framework that can be applied to virtually any source separation problem by simply selecting from a library of constraints suitable constraints accounting for the available information about that source (see Fig. 1, bottom). More precisely, we wish such a framework to be

- *general*, i.e., generalizing existing methods and making it possible to combine them,
- *flexible*, allowing easy incorporation of the *a priori* knowledge about a particular problem considered.

To achieve the property of generality, we need to find some common formulation for methods we would like to generalize. Many recently proposed methods for audio source separation and/or characterization [2]–[19] (see also [1] and references therein) are based on the same so-called *local Gaussian model* describing both the properties of the sources and of the mixing process. Thus, we chose this model as the core of our framework. To achieve flexibility, we fix the global structure of Gaussian covariances, and by means of a parametric model allow the introduction of knowledge about each individual source and its mixing characteristics via constraints on individual parameter subsets. The global

A. Ozerov and E. Vincent are with INRIA, Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes cedex, France (e-mails: alexey.ozerov@inria.fr, emmanuel.vincent@inria.fr).

F. Bimbot is with IRISA, CNRS - UMR 6074, Campus de Beaulieu, 35042 Rennes cedex, France (e-mail: frederic.bimbot@irisa.fr).

This work was partly supported by OSEO, the French State agency for innovation, under the Quaero program, and by the French Ministry of Foreign and European Affairs, the French Ministry of Higher Education and Research and the German Academic Exchange Service under project Procope 20142UD.

structure we consider corresponds to a generative model of the data that is motivated by the physics of the modeled processes, e.g., the source-filter model to represent a sound source and an approximation of the convolutive filter to represent its mixing characteristics. In summary, our framework generalizes the methods from [2]–[19], and, thanks to its flexibility, it becomes applicable in many other scenarios one can imagine.

We implement our framework using a generalized expectation-maximization (GEM) algorithm [20], where the M-step is solved by alternating optimization of different parameter subsets, taking the corresponding constraints into account and using multiplicative update (MU) rules inspired from the nonnegative matrix factorization (NMF) methodology (see, e.g., [9]) to update the nonnegative spectral parameters. Such an implementation is in fact possible thanks to the Gaussianity assumption leading to closed form update equations. The idea of mixing GEM algorithm with MU rules was already reported in [21] in the case of plain NMF spectral models and rank-1 spatial models, and we extend it here to the newly proposed structures. Our algorithmic contribution consists of (i) identifying the *GEM-MU* approach as suitable thanks to the implementability of the configurable framework, the simplicity of the update rules, the implicit verification of nonnegative constraints and its good convergence speed; and (ii) deriving of the update rules for the new model structures.

Our approach is in line with the *library of components* by Cardoso *et al* [22] developed for the separation of components in astrophysical images. However, we consider advanced audio-specific structures inspired by [1], [23] for source spectral power, as opposed to the unique block structure in [22] based on the assumption that source power is constant in some pre-defined region of time and space. In that sense, our framework is more flexible than [22]. Besides the framework itself, we propose a new structure for NMF-like decompositions of source power spectrograms, where the temporal envelope associated with each spectral pattern is represented as a nonnegative linear combination of time-localized temporal patterns. This structure can be used to ensure temporal continuity, but also to model more complex temporal characteristics, such as the attack or decay parts of a note. In line with time-localized patterns we include in our framework the so-called narrowband spectral patterns that allow constraining spectral patterns to be harmonic, inharmonic or noise-like. These structures were already reported in [14], [15], but only in case of harmonic constraints. Moreover, they were not applied for source separation so far. As compared to [24], where some preliminary aspects of this work were presented, we here present the framework in details, describe its implementation, and extend the experimental part illustrating the framework. Moreover, we propose an original mixing model formulation that allows the representation and the estimation of rank-1 [5] and full-rank [19] (actually any rank) spatial mixing models in a homogeneous way, thus enabling the combination of both models within a given mixture. Finally, we provide a proper probabilistic formulation of local Gaussian modeling for quadratic time-frequency representations [18] that supports and justifies the formulation given in [18].

We have also implemented and released a baseline version of the framework in Matlab. The corresponding software tool named *Flexible Audio Source Separation Toolbox (FASST)* is available at [25] together with a user guide, examples of usage (where the constraints are specified) and the corresponding audio examples. Given a source separation problem, one can choose one or few suitable constraint combinations based on his/her expertise and on the a priori knowledge, and then test all of them using FASST so as to select the best one.

In summary, the main contributions of this work include

- a general modeling structure,
- a general estimation algorithm,
- new spectral an temporal structures (time-localized patterns, narrowband spectral patterns),
- the implementation and distribution of a baseline version of the framework (the FASST toolbox [25]).

The rest of this paper is organized as follows. In Section II, existing approaches generalized by the proposed framework are discussed and an overview of the framework is given. Sections III and IV provide a detailed description of the framework and its algorithmic implementation. Thus, Section II is devoted to a reader interested in understanding the main principles of the framework and the physical meaning of the objects, and Sections III and IV to one willing to go deeper into the technical details. The results of a few source separation experiments are given in Section V to illustrate the flexibility of our framework and its potential performance improvement compared to individual approaches. Conclusions are drawn in Section VI.

II. Related existing approaches and framework overview

Source separation methods based on the local Gaussian model can be characterized by the following assumptions [1], [2], [5], [13], [19]:

1) *Gaussianity:* in some time-frequency (TF) representation the sources are modeled in each TF bin by zero-mean Gaussian random variables.
2) *Independence:* conditionally to their covariance matrices, these random variables are independent over time, frequency and between sources.
3) *Factorization of spectral and spatial characteristics:* for each TF bin, the covariance matrix of each source is expressed as the product of a *spatial covariance* matrix representing its spatial characteristics and a scalar *spectral power* representing its spectral characteristics.
4) *Linearity of mixing:* the mixing process translates into addition in the covariance domain.

A. State-of-the-art approaches based on the local Gaussian model

The state-of-the-art approaches [2]–[19] cover a wide range of source separation problems and models expressed via particular structures of local Gaussian covariances, including:

1) *Problem dimensionality:* Denoting by I and J, respectively, the number of channels of the observed

mixture and the number of sources to separate, the *single-channel* ($I = 1$) case is addressed in [6], and *underdetermined* ($1 < I < J$) and *(over-)determined* ($I \geq J$) cases are addressed in [5] and [2], respectively.

2) *Spatial covariance model: Instantaneous* and *convolutive* mixtures of point sources are modeled by *rank-1* spatial covariance matrices in [5] and [3], respectively. In [19] reverberant convolutive mixtures of point sources are modeled by *full-rank* spatial covariance matrices that, in contrast to rank-1 covariance matrices, can account for the spatial spread of each source induced by the reverberation.
3) *Spectral power model:* Several models were proposed for the spectral power, e.g., *unconstrained* models [10], *block constant* models [5], Gaussian mixture models (GMM) or hidden Markov models (HMM) [2], Gaussian scaled mixture models (GSMM) or scaled HMMs (S-HMM) [13], NMF [4] together with its variants, harmonic NMF [14] or temporal activation constrained NMF [9], and source-filter models [16]. These models are suitable for the representation of different types of sources, for example GSMM is rather suitable for a monophonic source, e.g., speech, and NMF for a polyphonic one, e.g., polyphonic musical instrument, [13].
4) *Input representation:* While the most of the considered methods use the short time Fourier transform (STFT) as the input TF representation, some of them, e.g., [14], [15], [18], use the auditory-motivated equivalent rectangular bandwidth (ERB) quadratic representation. More generally, we consider here both *linear representations*, where the signal is represented by a vector of complex-valued coefficients in each TF bin, as well as *quadratic representations*, where the signal is represented via its local covariance matrix in each TF bin [26].

Table I provides an overview of some of the local Gaussian model-based approaches considered here, where the specificities of each method are marked by crosses $\times$. We see from Table I that a few of these methods have already been combined together, for example GSMM and NMF were combined in [8], and NMF [9] was combined with rank-1 and full-rank mixing models in [13] and [17], respectively. However, many combinations have not yet been investigated. Indeed, assuming that each source follows one of the 3 spatial covariance models and one of the 8 spectral variance models from Table I, the total number of configurations equals to 2×24^J for J sources (in fact much more since each source can follow several spectral variance models at the same time), while Table I reports only 16 existing configurations.

B. Other related state-of-the-art approaches

While the local Gaussian model-based framework offers maximum of flexibility, there exist some methods that do not satisfy (fully or partially) the aforementioned assumptions and are thus not strictly covered by the framework. Nevertheless, our framework allows the implementation of similar structures. Let us give some examples. Binary masking-based source estimation [27], [28] does not satisfy the source independence assumption. However, it is known to perform poorly compared to local Gaussian model-based separation, as it was shown in [13], [18] for convolutive mixtures [1] and demonstrated through the signal separation evaluation campaigns SiSEC 2008 [30] and SiSEC 2010 [29], where for instantaneous mixtures local Gaussian model-based approaches gave better results than the *oracle* (using the ground truth) binary masks. The methods proposed in [31], [32] are also based on Gaussian models albeit in the time domain. Notably, time sample-based GMMs and time-varying autoregressive models are considered as source models in [31] and [32], respectively. However, the number of existing time-domain structures is fairly reduced. Our TF domain models make it possible to account for these structures by means of suitable constraints over spectral power, while allowing their combination with more advanced structures. There are also many works on NMF and its extensions [33]–[38] and on GMMs / HMMs [39], [40] based on nongaussian models of the complex-valued STFT coefficients. These models are essentially covered by our framework in the sense that we can implement similar or equivalent model structures, albeit under Gaussian assumptions. The benefit of local Gaussian modeling is that it naturally leads to closed-form expressions in the multichannel case and allows the modeling of diffuse sources [19], contrary to the models in [33]–[40]. Finally, according to Cardoso [41], nongaussianity and nonstationarity are alternative routes to source separation, such that nonstationary nongaussian models would offer little benefit compared to nonstationary Gaussian models in terms of separation performance despite considerably greater computation cost.

C. Framework overview

We now present an overview of the proposed framework focusing on the most important concepts. An exhaustive description is given in Sections III and IV.

The framework is based on a flexible model described by parameters $\theta = \{\theta_j\}_{j=1}^J$, where θ_j are the parameters of the j-th source ($j = 1, \ldots, J$). Each θ_j is split in turn into nine parameter subsets according to a fixed structure, as described below and summarized in Table II.

1) Model structure: The parameters of j-th source include a complex-valued tensor $\mathbf{A}_j$ modeling its spatial covariance, and eight nonnegative matrices $(\theta_{j,2}, \ldots, \theta_{j,9})$ modeling its spectral power over all TF bins.

The spectral power, denoted as $\mathbf{V}_j$, is assumed to be the product of an *excitation spectral power* $\mathbf{V}_j^{\mathrm{ex}}$, representing, e.g., the excitation of the glottal source for voice or the plucking of the string of a guitar, and a *filter spectral power* $\mathbf{V}_j^{\mathrm{ft}}$, representing, e.g., the vocal tract or the impedance of the guitar body [23], [35]. While such a model is usually called source-filter model, we call it here *excitation-filter model* in order to avoid possible confusions with the "sources" to be separated.

[1] Binary masking-based approaches can still be quite powerful for convolutive mixtures, as demonstrated in [29]. Thus, a good way to proceed is probably to use them to initialize local Gaussian model-based approaches, as it is done in [13], and as we do in the experimental part.

Reference		[7]	[6]	[8]	[16]	[4]	[14]	[15]	[9]	[5]	[11]	[13]	[19]	[18]	[17]	[3]	[2]
Problem dimensionality	single-channel	×	×	×	×	×	×	×									
	underdetermined									×	×	×	×	×	×		
	(over-)determined															×	×
Spatial covariance model	rank-1 instantaneous									×	×						
	rank-1 convolutive											×				×	×
	full-rank												×	×	×		
Spectral variance model	unconstrained												×	×			
	block constant									×						×	
	GMM / HMM	×									×						×
	GSMM / S-HMM		×	×	×												
	NMF			×	×	×						×			×		
	harmonic NMF						×	×									
	temp. constr. NMF							×	×								
	source-filter	×			×												
Input representation	linear	×	×	×	×	×			×	×	×	×	×		×	×	×
	quadratic						×	×						×			

TABLE I
SOME STATE-OF-THE-ART LOCAL GAUSSIAN MODEL-BASED APPROACHES FOR AUDIO SOURCE SEPARATION.

The excitation spectral power $\mathbf{V}_j^{\mathrm{ex}}$ is further decomposed as the sum of *characteristic spectral patterns* $\mathbf{E}_j^{\mathrm{ex}}$ modulated by *time activation coefficients* $\mathbf{P}_j^{\mathrm{ex}}$ [4], [9]. Each characteristic spectral pattern may be associated for instance with one specific pitch, so that the time activation coefficients denote which pitches are active on each time frame. In order to further constrain the fine structure of the spectral patterns, they are represented as linear combinations of *narrowband spectral patterns* $\mathbf{W}_j^{\mathrm{ex}}$ [14] with weights $\mathbf{U}_j^{\mathrm{ex}}$. These narrowband patterns may be for instance harmonic, inharmonic or noise-like and the weights determine the overall spectral envelope. Following the same idea, we propose here to represent the series of time activation coefficients $\mathbf{P}_j^{\mathrm{ex}}$ as sums of *time-localized patterns* $\mathbf{H}_j^{\mathrm{ex}}$ with weights $\mathbf{G}_j^{\mathrm{ex}}$. The time-localized patterns may represent the typical temporal shape of the notes while the weights encode their onset times. Different temporal fine structures such as continuity or specific rhythm patterns may also be accounted for in this way. Note that temporal models of the activation coefficients have been proposed in the state-of-the-art, using probabilistic priors [9], [34], note-specific Gaussian-shaped time-localized patterns [42], or unstructured TF patterns [33]. Our proposition is complementary to [9], [34] in that it accounts for temporal behaviour in the model structure itself in addition to possible priors on the model parameters. Moreover, it is more flexible than [9], [34], [42], since it allows the modeling of other characteristics than continuity or sparsity. Finally, while it can model similar TF patterns to [33], it involves much fewer parameters, which typically leads to more robust parameter estimation.

The filter spectral power $\mathbf{V}_j^{\mathrm{ft}}$ is similarly expressed in terms of characteristic spectral patterns $\mathbf{E}_j^{\mathrm{ft}}$ modulated by time activation coefficients [16], which are in turn decomposed into narrowband spectral patterns $\mathbf{W}_j^{\mathrm{ft}}$ with weights $\mathbf{U}_j^{\mathrm{ft}}$ and time-localized patterns $\mathbf{H}_j^{\mathrm{ft}}$ with weights $\mathbf{G}_j^{\mathrm{ft}}$, respectively. In the case of speech or singing voice, each characteristic spectral pattern may represent the spectral formants of a given phoneme, while the plosiveness and the sequence of pronounced phonemes may be encoded by the time-localized patterns and the associated weights.

In summary, as it will be explained in details in Section III-E, the spectral power of each source obeys a three-level hierarchical nonnegative matrix decomposition structure (see equations (9), (10), (12), (13) and Figures 3 and 4 below) including at the bottom level the eight parameter subsets $\mathbf{W}_j^{\mathrm{ex}}$, $\mathbf{U}_j^{\mathrm{ex}}$, $\mathbf{G}_j^{\mathrm{ex}}$, $\mathbf{H}_j^{\mathrm{ex}}$, $\mathbf{W}_j^{\mathrm{ft}}$, $\mathbf{U}_j^{\mathrm{ft}}$, $\mathbf{G}_j^{\mathrm{ft}}$ and $\mathbf{H}_j^{\mathrm{ft}}$ (see Eq. (13)).

Parameter subsets		Size	Range
$\theta_{j,1} = \mathbf{A}_j$	mixing parameters	$I \times R_j \times F \times N$	$\in \mathbb{C}$
$\theta_{j,2} = \mathbf{W}_j^{\mathrm{ex}}$	ex. narrowband spectral patterns	$F \times L_j^{\mathrm{ex}}$	$\in \mathbb{R}_+$
$\theta_{j,3} = \mathbf{U}_j^{\mathrm{ex}}$	ex. spectral pattern weights	$L_j^{\mathrm{ex}} \times K_j^{\mathrm{ex}}$	$\in \mathbb{R}_+$
$\theta_{j,4} = \mathbf{G}_j^{\mathrm{ex}}$	ex. time pattern weights	$K_j^{\mathrm{ex}} \times M_j^{\mathrm{ex}}$	$\in \mathbb{R}_+$
$\theta_{j,5} = \mathbf{H}_j^{\mathrm{ex}}$	ex. time-localized patterns	$M_j^{\mathrm{ex}} \times N$	$\in \mathbb{R}_+$
$\theta_{j,6} = \mathbf{W}_j^{\mathrm{ft}}$	ft. narrowband spectral patterns	$F \times L_j^{\mathrm{ft}}$	$\in \mathbb{R}_+$
$\theta_{j,7} = \mathbf{U}_j^{\mathrm{ft}}$	ft. spectral pattern weights	$L_j^{\mathrm{ft}} \times K_j^{\mathrm{ft}}$	$\in \mathbb{R}_+$
$\theta_{j,8} = \mathbf{G}_j^{\mathrm{ft}}$	ft. time pattern weights	$K_j^{\mathrm{ft}} \times M_j^{\mathrm{ft}}$	$\in \mathbb{R}_+$
$\theta_{j,9} = \mathbf{H}_j^{\mathrm{ft}}$	ft. time-localized patterns	$M_j^{\mathrm{ft}} \times N$	$\in \mathbb{R}_+$

TABLE II
PARAMETER SUBSETS $\theta_{j,k}$ ($j = 1, \ldots, J$, $k = 1, \ldots, 9$) ENCODING THE STRUCTURE OF EACH SOURCE.

2) Constraints: Given the above fixed model structure, prior information about each source can now be exploited by specifying deterministic or probabilistic constraints over each parameter subset of Table II. Examples of such constraints are given in Table III. Each parameter subset can be fixed [2] (i.e., unchanged during estimation), adaptive (i.e., fully fitted to the mixture) or partially adaptive (only some parameters within the subset are adaptive). In the latter two cases, a probabilistic prior, such as a continuity prior [9] or a sparsity-inducing prior [4], can be specified over the parameters. The mixing parameters $\mathbf{A}_j$ can be time-varying or time-invariant (in Table III the latter case is only considered), frequency-dependent for convolutive mixtures or frequency-independent for instantaneous mixtures. Mixing parameters $\mathbf{A}_j$ can be given a probabilistic prior as well. E.g., it can be a Gaussian prior with the mean corresponding to the parameters of a presumed direction and with the covariance matrix representing

[2]The fixed parameters can be either set manually or learned beforehand from some training data. Learning is equivalent to model parameter estimation over the training data and can thus be achieved using our framework.

a degree of uncertainty about this direction. The rank R_j ($1 \leq R_j \leq I$) of the spatial covariance is specifiable via the size of tensor $\mathbf{A}_j$ (see Table II). Each parameter subset may also be constrained to have a limited number of nonzero entries. For instance, every column of $\mathbf{G}_j^{\text{ex}}$ and / or $\mathbf{G}_j^{\text{ft}}$ may be constrained to have a single nonzero entry accounting for a GSMM / S-HMM structure or a single nonzero entry equal to 1 accounting for a GMM / HMM structure.

Parameter subsets	Constraint	Value
$\mathbf{A}_j, \mathbf{W}_j^{\text{ex}}, \mathbf{U}_j^{\text{ex}}, \mathbf{G}_j^{\text{ex}}, \mathbf{H}_j^{\text{ex}}, \mathbf{W}_j^{\text{ft}}, \mathbf{U}_j^{\text{ft}}, \mathbf{G}_j^{\text{ft}}, \mathbf{H}_j^{\text{ft}}$	degree of adaptability	'fixed'
		'part_adapt'
		'adapt'
$\mathbf{A}_j$	mixing stationarity	'time_inv'
	mixing type	'conv'
		'inst'
$\mathbf{G}_j^{\text{ex}}, \mathbf{G}_j^{\text{ft}}$	temporal constraint	'null'
		'GMM', 'HMM'
		'GSMM', 'SHMM'

TABLE III
EXAMPLES OF USER-SPECIFIABLE CONSTRAINTS OVER THE PARAMETER SUBSETS.

3) Estimation algorithm: Given the above model structure and constraints, source separation can be achieved in two steps as shown in Fig. 2. First, given initial parameter values, the model parameters θ are estimated from the mixture $\mathbf{X}$ using an iterative GEM algorithm, where the E-step consists in computing some quantity $\widehat{\mathbb{T}}$ called *conditional expectation of the natural statistics*, and the M-step consists in updating the parameters θ given $\widehat{\mathbb{T}}$ by alternating optimization of each of the $J \times 9$ parameter subsets. This allows taking any combination of constraints specified by user into account. Second, given the mixture $\mathbf{X}$ and the estimated model parameters θ, source estimates $\widehat{\mathbf{Y}}$ are computed using Wiener filtering.

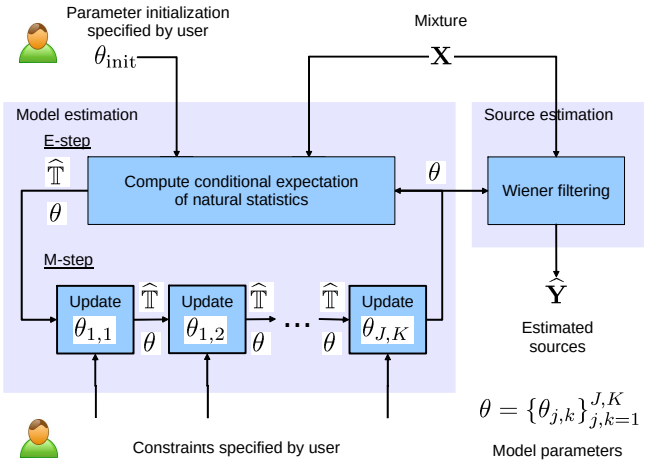

Fig. 2. Overview of the proposed general algorithm for parameter estimation and source separation.

D. FASST toolbox: Current baseline implementation

The FASST toolbox (released and available at [25]) implements so far a baseline version of the framework in Matlab that covers only the library of constraints summarized in Table III for mono or stereo recordings ($I = 1$ or $I = 2$). This restriction to up to $I = 2$ channels enables the use of a 2×2 matrix inversion trick described in [13] that leads to an efficient implementation in Matlab. However, the framework itself is neither restricted to the constraints in Table III nor to mono / stereo mixtures.

III. DETAILED STRUCTURE AND EXAMPLE CONSTRAINTS

In this section we describe in details the nine parameter subsets modeling each source and some example constraints. We also introduce the detailed notations to be used in the rest of the paper.

A. Formulation of the audio source separation problem

We assume that the observed I-channel time-domain signal, called *mixture*, $\tilde{\mathbf{x}}(t) \in \mathbb{R}^I$, $t = 1, \ldots, T$, is the sum of J multichannel signals $\tilde{\mathbf{y}}_j(t) \in \mathbb{R}^I$, called *spatial source images* [1], [22]:

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

The goal of source separation is to estimate the spatial source images $\tilde{\mathbf{y}}_j(t)$ given the mixture $\tilde{\mathbf{x}}(t)$. This now common formulation is more general than the convolutive formulation in [13], which is restricted to point sources [1], [22].

B. Input representation

Audio signals are usually processed in the TF domain, due to their sparsity in this domain. Two families of input representations are considered in the literature, namely *linear* [13] and *quadratic* [18] representations.

1) Linear representations: After applying a linear complex-valued TF transform, the mixture (1) becomes:

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{j,fn}, \tag{2}$$

where $\mathbf{x}_{fn} \in \mathbb{C}^I$ and $\mathbf{y}_{j,fn} \in \mathbb{C}^I$ are I-dimensional complex-valued vectors of TF coefficients of the corresponding time-domain signals; and $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote respectively frequency bin and time-frame index. This formulation covers the STFT, that is the most popular TF representation used for audio source separation.

2) Quadratic representations: A few studies have relied on quadratic representations instead, where the signal is described in each TF bin by its empirical $I \times I$ covariance matrix [5], [10], [18]

$$\widehat{\mathbf{R}}_{\mathbf{x},fn} = \widehat{\mathbb{E}}[\mathbf{x}_{fn}\mathbf{x}_{fn}^H], \tag{3}$$

where $\widehat{\mathbb{E}}[\cdot]$ denotes *empirical expectation* computed, e.g., by local averaging of the STFT [5], [10] or of the input of an ERB filterbank [18]. Note that linear representations are special cases of quadratic representations with $\widehat{\mathbf{R}}_{\mathbf{x},fn} = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$. Quadratic representations include additional information about the local correlation between channels which often increases the accuracy of parameters estimation [10]. In the following, we use the linear notations $\mathbf{x}_{fn}$ and $\mathbf{y}_{j,fn}$ for simplicity and include the empirical expectation when appropriate. A more rigorous derivation of the local Gaussian model for quadratic representations is given in Appendix A.

C. Local Gaussian model

We assume that in each TF bin, each source $\mathbf{y}_{j,fn} \in \mathbb{C}^I$ is a proper complex-valued Gaussian random vector with zero mean and covariance matrix $\mathbf{\Sigma}_{\mathbf{y},j,fn} = v_{j,fn}\mathbf{R}_{j,fn}$

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_c\left(\bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right), \tag{4}$$

where the matrix $\mathbf{R}_{j,fn} \in \mathbb{C}^{I\times I}$ called *spatial covariance matrix* represents the spatial characteristics of the source and of the mixing setup, and the non-negative scalar $v_{j,fn} \in \mathbb{R}_+$ called *spectral power* represents the spectral characteristics of the source [1]. Moreover, the random vectors $\mathbf{y}_{j,fn}$ are assumed to be mutually independent given $\mathbf{\Sigma}_{\mathbf{y},j,fn}$.

D. Spatial covariance structure and example constraints

1) Structure: In the case of audio, it is mostly interesting to consider either rank-1 spatial covariances representing instantaneously or convolutively mixed point sources with low reverberation [13] or full-rank spatial covariances modeling diffuse or reverberated sources [19]. More generally, we assume covariances of any positive rank. Let $0 < R_j \leq I$ be the rank of covariance $\mathbf{R}_{j,fn}$. This matrix can then be non-uniquely represented as [3]

$$\mathbf{R}_{j,fn} = \mathbf{A}_{j,fn}\mathbf{A}_{j,fn}^H, \tag{5}$$

where $\mathbf{A}_{j,fn}$ is an $I \times R_j$ complex-valued matrix of rank R_j. Moreover, for every source j and for every TF bin (f,n) we introduce R_j independent Gaussian random variables $s_{jr,fn}$ $(r = 1,\ldots,R_j)$ distributed as

$$s_{jr,fn} \sim \mathcal{N}_c\left(0, v_{j,fn}\right). \tag{6}$$

With these notations the model defined by (2) and (4) is equivalent to the following mixture of $R = \sum_{j=1}^{J} R_j$ point *sub-sources* $s_{jr,fn}$:

$$\mathbf{x}_{fn} = \mathbf{A}_{fn}\mathbf{s}_{fn}, \tag{7}$$

where $\mathbf{s}_{fn} = [\mathbf{s}_{1,fn}^T,\ldots,\mathbf{s}_{J,fn}^T]^T$ is an $R \times 1$ vector of sub-source coefficients with $\mathbf{s}_{j,fn} = [s_{j1,fn},\ldots,s_{jR_j,fn}]^T$, and $\mathbf{A}_{fn} = [\mathbf{A}_{1,fn},\ldots,\mathbf{A}_{J,fn}]$ is an $I \times R$ mixing matrix. Thus, for a given TF bin (f,n) our model is equivalent to a complex-valued linear mixture of R sub-sources (7), where the sub-sources $s_{jr,fn}$ $(r = 1,\ldots,R_j)$ associated with the same source j share the same spectral power (6). We suppose that the rank R_j is specified for every source j.

2) Example constraints: In our baseline implementation we assume that the spatial covariances are time-invariant, i.e., $\mathbf{A}_{j,fn} = \mathbf{A}_{j,f}$. Moreover, we assume that for every source j the spatial parameters $\mathbf{A}_j$ can be either instantaneous (i.e., constant over frequency and real-valued: $\mathbf{A}_{j,fn} = \mathbf{A}_{j,n} \in \mathbb{R}^{I\times R_j}$) or convolutive (i.e., frequency-dependent), and either fixed, adaptive or partially adaptive. Some examples of constraints are given in Table III.

[3]Such an R_j-rank covariance matrix parametrization was inspired by [22], where $\mathbf{R}_{j,fn}$, intended to model correlated or multi-dimensional components, is parametrized as $\mathbf{R}_{j,fn} = \mathbf{A}_{j,fn}\mathbf{P}_{j,fn}\mathbf{A}_{j,fn}^H$, where $\mathbf{P}_{j,fn}$ is a full-rank $R_j \times R_j$ positive matrix. However, our parametrization (5) is less redundant and it is applied for audio source separation, and not for separation of components in astrophysical images, as in [22].

E. Spectral power structure and example constraints

To model spectral power we use nonnegative hierarchical audio-specific decompositions [23], thus all variables introduced in this section are assumed to be non-negative.

1) Excitation-filter model: We first model the spectral power $v_{j,fn}$ as the product of an excitation spectral power $v_{j,fn}^{\mathrm{ex}}$ and a filter spectral power $v_{j,fn}^{\mathrm{ft}}$ [23], [35]:

$$v_{j,fn} = v_{j,fn}^{\mathrm{ex}} \times v_{j,fn}^{\mathrm{ft}}, \tag{8}$$

that can be rewritten as

$$\mathbf{V}_j = \mathbf{V}_j^{\mathrm{ex}} \odot \mathbf{V}_j^{\mathrm{ft}}, \tag{9}$$

where $\odot$ denotes element-wise matrix multiplication and $\mathbf{V}_j \triangleq [v_{j,fn}]_{f,n}$, $\mathbf{V}_j^{\mathrm{ex}} \triangleq [v_{j,fn}^{\mathrm{ex}}]_{f,n}$, $\mathbf{V}_j^{\mathrm{ft}} \triangleq [v_{j,fn}^{\mathrm{ft}}]_{f,n}$.

Figure 3 gives an example of the excitation-filter decomposition (9) as applied to the spectral power of several guitar notes. In this example the filter $\mathbf{V}_j^{\mathrm{ft}}$ is time-invariant with lowpass characteristics, and the excitation $\mathbf{V}_j^{\mathrm{ex}}$ is a time-varying combination of few characteristic spectral patterns. However, in the most of realistic situations both the excitation and the filter are time-varying. Thus, the excitation-filter model with time-varying excitation and filter is a physically-motivated generative model that is suitable for many audio sources. While time-invariant filters were considered, e.g., in [7], [35], some approaches consider time-varying filters [16], [43]. We believe that our framework opens a door for further investigation of time-varying filters.

2) Excitation power structure: The excitation spectral power $[v_{j,fn}^{\mathrm{ex}}]_f$ is modeled as the sum of K_j^{ex} characteristic spectral patterns $[e_{j,fk}^{\mathrm{ex}}]_f$ modulated in time by $p_{j,kn}^{\mathrm{ex}}$, i.e., $v_{j,fn}^{\mathrm{ex}} = \sum_{k=1}^{K_j^{\mathrm{ex}}} p_{j,kn}^{\mathrm{ex}} e_{j,fk}^{\mathrm{ex}}$ [9]. Introducing the matrices $\mathbf{P}_j \triangleq [p_{j,kn}^{\mathrm{ex}}]_{k,n}$ and $\mathbf{E}_j^{\mathrm{ex}} \triangleq [e_{j,fk}^{\mathrm{ex}}]_{f,k}$ it can be rewritten as

$$\mathbf{V}_j^{\mathrm{ex}} = \mathbf{E}_j^{\mathrm{ex}}\,\mathbf{P}_j^{\mathrm{ex}}. \tag{10}$$

In order to further constrain the spectral fine structure of the spectral patterns, they are represented as linear combinations of L_j^{ex} narrowband spectral patterns $[w_{j,fl}^{\mathrm{ex}}]_f$ [14], i.e., $e_{j,fk}^{\mathrm{ex}} = \sum_{l=1}^{L_j^{\mathrm{ex}}} u_{j,lk}^{\mathrm{ex}} w_{j,fl}^{\mathrm{ex}}$, where $u_{j,lk}^{\mathrm{ex}}$ are non-negative weights. The series of time activation coefficients $p_{j,kn}^{\mathrm{ex}}$ are also represented as sums of M_j^{ex} time-localized patterns, i.e., $p_{j,kn}^{\mathrm{ex}} = \sum_{m=1}^{M_j^{\mathrm{ex}}} h_{j,mn}^{\mathrm{ex}} g_{j,km}^{\mathrm{ex}}$. Altogether we have:

$$v_{j,fn}^{\mathrm{ex}} = \sum\nolimits_{k=1}^{K_j^{\mathrm{ex}}} \sum\nolimits_{m=1}^{M_j^{\mathrm{ex}}} h_{j,mn}^{\mathrm{ex}} g_{j,km}^{\mathrm{ex}} \sum\nolimits_{l=1}^{L_j^{\mathrm{ex}}} u_{j,lk}^{\mathrm{ex}} w_{j,fl}^{\mathrm{ex}}, \tag{11}$$

and, introducing matrices $\mathbf{H}_j^{\mathrm{ex}} \triangleq [h_{j,mn}^{\mathrm{ex}}]_{m,n}$, $\mathbf{G}_j^{\mathrm{ex}} \triangleq [g_{j,km}^{\mathrm{ex}}]_{k,m}$, $\mathbf{U}_j^{\mathrm{ex}} \triangleq [u_{j,lk}^{\mathrm{ex}}]_{l,k}$ and $\mathbf{W}_j^{\mathrm{ex}} \triangleq [w_{j,fl}^{\mathrm{ex}}]_{f,l}$, this equation can be rewritten in matrix form as

$$\mathbf{V}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}}\,\mathbf{U}_j^{\mathrm{ex}}\,\mathbf{G}_j^{\mathrm{ex}}\,\mathbf{H}_j^{\mathrm{ex}}. \tag{12}$$

Figure 4 shows an example of the excitation structure $\mathbf{V}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}}\,\mathbf{U}_j^{\mathrm{ex}}\,\mathbf{G}_j^{\mathrm{ex}}\,\mathbf{H}_j^{\mathrm{ex}}$, as applied to six notes played on a xylophone. In this example, the narrowband spectral patterns $\mathbf{W}_j^{\mathrm{ex}}$ include 66 harmonic patterns modeling the harmonic part of 11 notes and 9 smooth patterns modeling the attacks, and the matrix of weights $\mathbf{U}_j^{\mathrm{ex}}$ is very sparse so as to eliminate invalid combinations of narrowband spectral patterns (e.g., a

characteristic spectral pattern should not be a combination of narrowband spectral patterns with different pitches). The time-localized patterns $\mathbf{H}_j^{\text{ex}}$ include decreasing exponentials to model the decay part of the notes and discrete Dirac functions to model note attacks, and the matrix of weights $\mathbf{G}_j^{\text{ex}}$ is sparse so as not to allow the attacks (smooth spectral patterns) to be modulated by exponential temporal patterns and not to allow harmonic note parts (harmonic spectral patterns) to be modulated by Dirac temporal patterns. Such a structure is a simplified version of the conventional attack-decay-sustain-release model (see, e.g., [44]). More sophisticated structures, where, e.g., the sustain and release parts are modeled by exponentials with different decrease rates can be implemented as well within our framework.

3) Filter power structure: The filter spectral power $[v_{j,fn}^{\text{ft}}]_f$ is represented using exactly the same structure as in (11).

4) Total power structure: Altogether the spectral power structure can be represented by the following nonnegative matrix decomposition (see also Table II)

$$\mathbf{V}_j = \left(\mathbf{W}_j^{\text{ex}}\,\mathbf{U}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}\right) \odot \left(\mathbf{W}_j^{\text{ft}}\,\mathbf{U}_j^{\text{ft}}\,\mathbf{G}_j^{\text{ft}}\,\mathbf{H}_j^{\text{ft}}\right). \quad (13)$$

Each matrix in this decomposition is subject to specific constraints presented below.

5) Example constraints: Each matrix $\theta_{j,k}$ ($k = 2, \ldots, 9$) in (13) can be fixed, adaptive or partially fixed (see Tab. III). In the latter two cases, a probabilistic prior $p(\theta_{j,k}|\eta_{j,k})$, such as a time continuity prior [9] or a sparsity-inducing prior [4] can be set. We denote by $\eta_{j,k}$ the *hyperparameters* of the prior that can be fixed or adaptive as well.

To cover *discrete state-based models* such as GMM, HMM, and their scaled versions GSMM, S-HMM, every column $\mathbf{g}_{j,m}^{\text{ex}} = [g_{j,km}^{\text{ex}}]_k$ of matrix $\mathbf{G}_j^{\text{ex}}$ (and similarly for matrix $\mathbf{G}_j^{\text{ft}}$) may further be constrained to have either a single nonzero entry (for GSMM, S-HMM) or a single nonzero entry equal to 1 (for GMM, HMM). Let $q_{j,m}^{\text{ex}} \in \{1, \ldots, K_j^{\text{ex}}\}$ be the index of the corresponding nonzero entry and $\mathbf{q}_j^{\text{ex}} = [q_{j,m}^{\text{ex}}]_m$ the resulting *state sequence* [4]. The prior distribution of $\theta_{j,4} = \mathbf{G}_j^{\text{ex}}$ with hyperparameters $\eta_{j,4} = \mathbf{\Lambda}_j^{\text{ex}}$ is defined as

$$p(\theta_{j,4}|\eta_{j,4}) = p(\mathbf{q}_j^{\text{ex}}|\mathbf{\Lambda}_j^{\text{ex}}) = \prod\nolimits_{m=2}^{M_j^{\text{ex}}} \lambda_{j,q_{j,m-1}^{\text{ex}} q_{j,m}^{\text{ex}}}^{\text{ex}}, \quad (14)$$

where $\mathbf{\Lambda}_j^{\text{ex}} = [\lambda_{j,kk'}^{\text{ex}}]_{k,k'}$ ($\lambda_{j,kk'}^{\text{ex}} = \mathbb{P}(q_{j,m}^{\text{ex}} = k'|q_{j,m-1}^{\text{ex}} = k)$) denotes the $K_j^{\text{ex}} \times K_j^{\text{ex}}$ state transition probability matrix with $\lambda_{j,kk'}^{\text{ex}}$ being independent on k (i.e., $\lambda_{j,kk'}^{\text{ex}} = \lambda_{j,k'}^{\text{ex}}$) in the case of GMM or GSMM. As discussed in [12], the discrete state-based models are rather suitable for monophonic sources (e.g., singing voice or wind instruments), while the unconstrained NMF decompositions are more appropriate for polyphonic sources (e.g., piano or guitar).

F. Generality

It can be easily shown that the model structures considered in [2]–[19] are particular instances of the proposed general formulation. Let us give some examples.

Pham *et al* [3] assume rank-1 spatial covariances and constant spectral power over time-frequency regions of size 1 frequency bin $\times$ L frames. This structure can be implemented in our framework by choosing rank-1 adaptive spatial time-invariant covariances, i.e., $\mathbf{A}_j$ is an adaptive tensor of size $2 \times 1 \times F \times N$ subject to the time-invariance constraint, and constraining the spectral power to $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}$ [5] with $\mathbf{W}_j^{\text{ex}}$ being the $F \times F$ identity matrix, $\mathbf{G}^{\text{ex}}$ a $F \times \lceil N/L \rceil$ adaptive matrix, and $\mathbf{H}_j^{\text{ex}}$ the $\lceil N/L \rceil \times N$ fixed matrix with entries $h_{j,mn}^{\text{ex}} = 1$ for $n \in \mathcal{L}_m$ and $h_{j,mn}^{\text{ex}} = 0$ for $n \notin \mathcal{L}_m$, where $\mathcal{L}_m$ is the set of time frames belonging to the m-th block.

Multichannel NMF structures with point source (rank-1) [13] or diffuse source (full-rank) [17] models can be represented within our framework as $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}$ [5] with $\mathbf{W}_j^{\text{ex}}$ and $\mathbf{G}_j^{\text{ex}}$ being adaptive matrices of size $F \times K_j^{\text{ex}}$ and $K_j^{\text{ex}} \times N$, respectively, and $\mathbf{A}_j$ being an adaptive tensor of size $2 \times 1 \times F \times N$ or $2 \times 2 \times F \times N$, respectively, subject to the time-invariance constraint.

Excitation-filter model-based separation of the main melody vs. the background music from single-channel recordings by Durrieu *et al.* [16] can be represented within our framework as follows. Mixing parameters $\mathbf{A}_j$ ($j = 1, 2$) are assumed to form a tensor of size $1 \times 1 \times F \times N$ with all the entries fixed to 1. The background music spectral power $\mathbf{V}_1$ is modeled exactly as in the case of the multichannel NMF described in the previous paragraph. The main melody spectral power is constrained to $\mathbf{V}_2 = (\mathbf{W}_2^{\text{ex}}\,\mathbf{G}_2^{\text{ex}}) \odot (\mathbf{W}_2^{\text{ft}}\,\mathbf{G}_2^{\text{ft}})$ [5] with $\mathbf{W}_2^{\text{ex}}$ being fixed and $\mathbf{G}_2^{\text{ex}}$, $\mathbf{W}_2^{\text{ft}}$ and $\mathbf{G}_2^{\text{ft}}$ being adaptive. Without any supplementary constraints this model is equivalent to the model referred as *instantaneous mixture model* in [16], and applying GSMM constraints to both the matrices $\mathbf{G}_2^{\text{ex}}$ and $\mathbf{G}_2^{\text{ft}}$ this model is equivalent to the model referred as *GSMM* in [16].

IV. Estimation algorithm

In this section we describe in details the proposed algorithm for the estimation of the model parameters and subsequent source separation.

A. Model estimation criterion

To estimate the model parameters, we use the standard maximum *a posteriori* (MAP) where the log-likelihood $\log p(\mathbf{x}_{fn}|\theta)$ in every TF point is replaced by its empirical expectation $\widehat{\mathbb{E}}[\log p(\mathbf{x}_{fn}|\theta)]$ according to the empirical expectation operator $\widehat{\mathbb{E}}[\cdot]$ introduced in Section III-B2 [10], [18]. Mathematically rigorous derivation of this criterion is given in Appendix A. This criterion consists in maximizing the *modified log-posterior* $\widehat{\mathcal{L}}(\theta, \eta|\mathbf{X}) \triangleq \widehat{\mathbb{E}}[\log p(\theta, \eta|\mathbf{X})]$, where $\mathbf{X} = \{\mathbf{x}_{fn}\}_{f,n}$, over the model parameters θ and the hyperparameters $\eta = \{\eta_{j,k}\}_{j,k=1}^{J,9}$. This quantity can be rewritten,

[4]Note that we consider here the state sequence $\mathbf{q}_j^{\text{ex}}$ as a parameter to be estimated, and not as a latent variable one integrates over, as it is usually done for GMM / HMM parameter estimation. This is indeed to achieve the goal of generality by making the E-step of the GEM algorithm independent of the specified constraints.

[5]Note that any set of matrices can be virtually removed from the spectral power decomposition (13). For example, one can obtain $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\,\mathbf{G}_j^{\text{ex}}\,\mathbf{H}_j^{\text{ex}}$ by assuming that the matrices $\mathbf{W}_j^{\text{ft}}$, $\mathbf{U}_j^{\text{ft}}$, $\mathbf{G}_j^{\text{ft}}$ and $\mathbf{H}_j^{\text{ft}}$ are of sizes $F \times 1$, 1×1, 1×1, and $1 \times N$, and that all their entries are fixed to 1, and that $\mathbf{U}_j^{\text{ex}} = \mathbf{I}_{K_j^{\text{ex}}}$ is the $K_j^{\text{ex}} \times K_j^{\text{ex}}$ identity matrix.

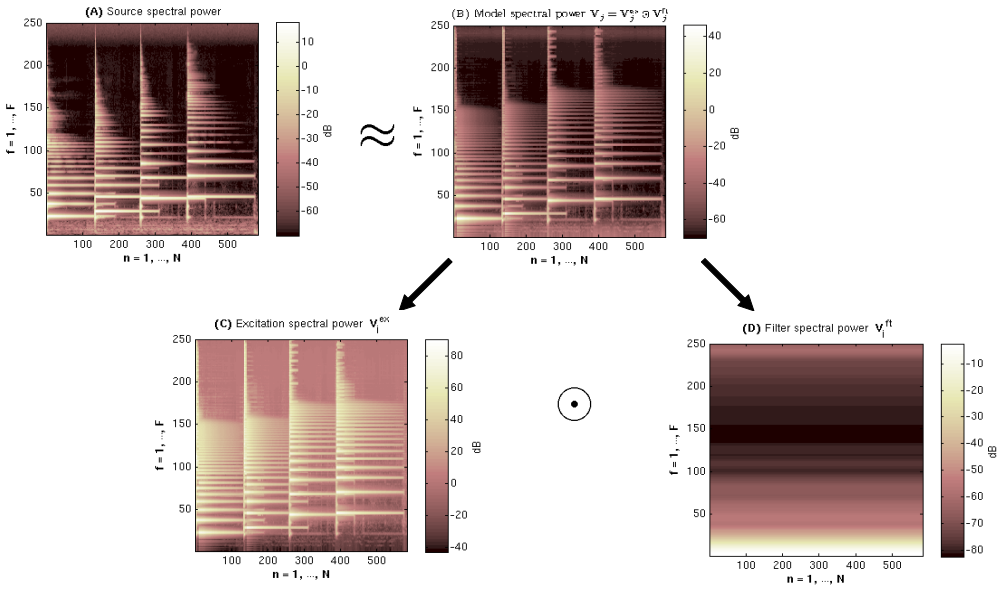

Fig. 3. Excitation-filter decomposition as applied to the spectral power of several guitar notes. **(A)**: source spectral power, **(B)**: model spectral power $\mathbf{V}_j = \mathbf{V}_j^{\mathrm{ex}} \odot \mathbf{V}_j^{\mathrm{ft}}$, **(C)**: excitation spectral power $\mathbf{V}_j^{\mathrm{ex}}$, **(D)**: filter spectral power $\mathbf{V}_j^{\mathrm{ft}}$.

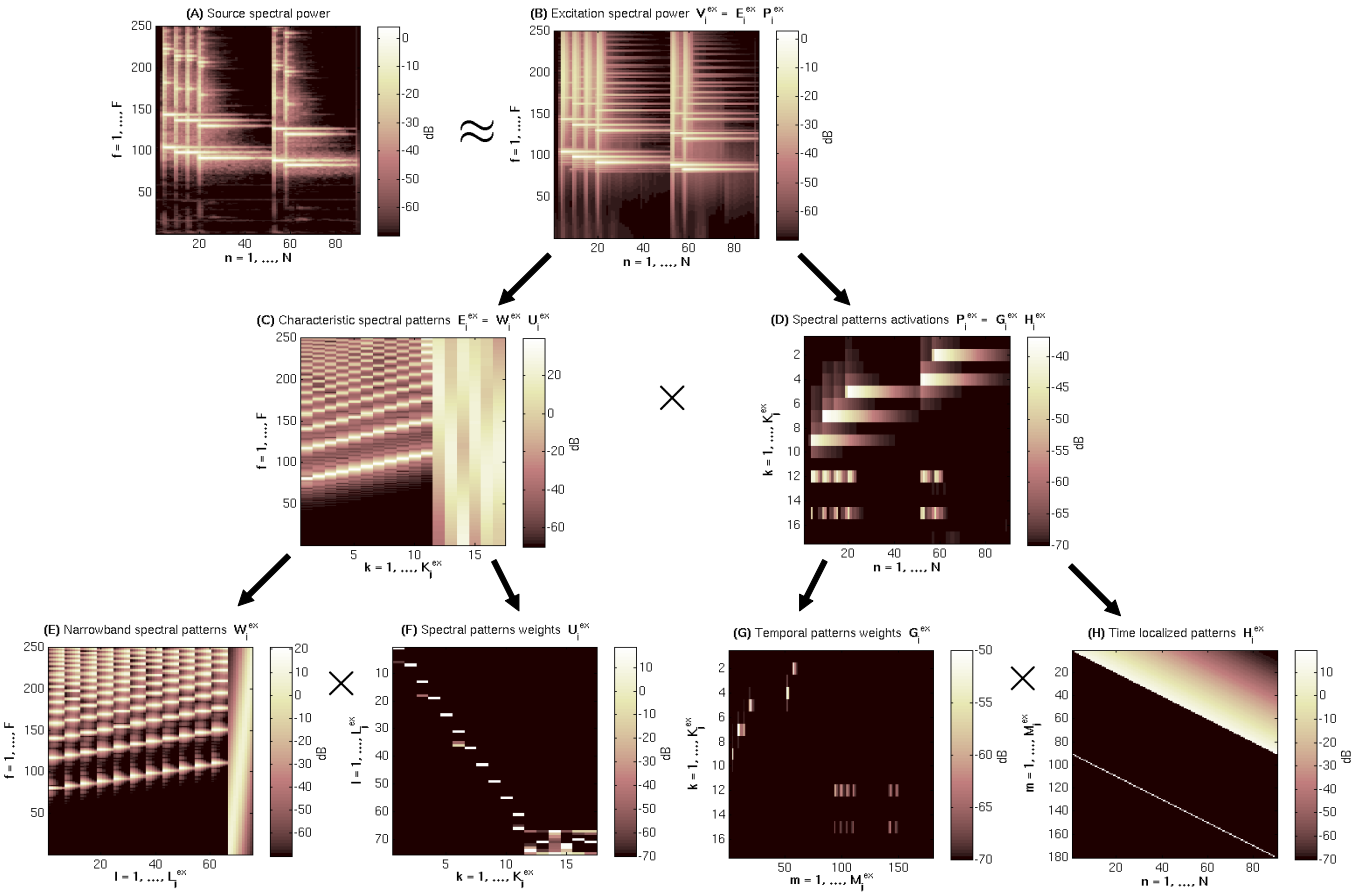

Fig. 4. Excitation power decomposition $\mathbf{V}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}} \mathbf{U}_j^{\mathrm{ex}} \mathbf{G}_j^{\mathrm{ex}} \mathbf{H}_j^{\mathrm{ex}}$ as applied to the spectral power of several xylophone notes. **(A)**: source spectral power, **(B)**: excitation spectral power $\mathbf{V}_j^{\mathrm{ex}} = \mathbf{E}_j^{\mathrm{ex}} \mathbf{P}_j^{\mathrm{ex}}$, **(C)**: characteristic spectral patterns $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}} \mathbf{U}_j^{\mathrm{ex}}$, **(D)**: spectral pattern activations $\mathbf{P}_j^{\mathrm{ex}} = \mathbf{G}_j^{\mathrm{ex}} \mathbf{H}_j^{\mathrm{ex}}$, **(E)**: narrowband spectral patterns $\mathbf{W}_j^{\mathrm{ex}}$, **(F)**: spectral pattern weights $\mathbf{U}_j^{\mathrm{ex}}$, **(G)**: temporal pattern weights $\mathbf{G}_j^{\mathrm{ex}}$, **(H)**: time-localized patterns $\mathbf{H}_j^{\mathrm{ex}}$.

using (2) and (4), as:

$$\widehat{\mathcal{L}}(\theta,\eta|\mathbf{X}) \stackrel{\text{c}}{=} \widehat{\mathcal{L}}(\mathbf{X}|\theta) + \log p(\theta|\eta) = \sum\nolimits_{f,n} \widehat{\mathbb{E}}[\log N_c(\mathbf{x}_{fn}|0,\mathbf{\Sigma}_{\mathbf{x},fn})] + \log p(\theta|\eta), \quad (15)$$

where $\mathbf{\Sigma}_{\mathbf{x},fn} \triangleq \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$, $\widehat{\mathcal{L}}(\mathbf{X}|\theta) \triangleq \widehat{\mathbb{E}}[\log p(\mathbf{X}|\theta)]$ is the *modified log-likelihood* and "$\stackrel{\text{c}}{=}$" denotes equality up to a constant. Using (3), the resulting criterion can be expressed as [13], [18]:

$$\theta^*,\eta^* = \arg\min_{\theta,\eta} \sum\nolimits_{f,n} \left[\text{tr}\left(\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}\widehat{\mathbf{R}}_{\mathbf{x},fn}\right) + \log|\mathbf{\Sigma}_{\mathbf{x},fn}|\right] - \sum\nolimits_{j,k=1}^{J,9} \log p(\theta_{j,k}|\eta_{j,k}). \quad (16)$$

We see that this criterion does not rely any more on the linear mixture representation $\mathbf{X}$, but only on the resulting empirical mixture covariances $\{\widehat{\mathbf{R}}_{\mathbf{x},fn}\}_{f,n}$.

B. Model estimation via a GEM algorithm

Given the model parameters $\theta = \{\theta_{j,k}\}_{j,k=1}^{J,9}$ specified in Table II and the hyperparameters $\eta = \{\eta_{j,k}\}_{j,k=1}^{J,9}$ together with user-defined constraints and initial values, we minimize the criterion (16) using a GEM algorithm [20] that consists in iterating the following expectation (E) and maximization (M) steps (see Fig. 2):

- *E-step:* Compute the conditional expectation of the so-called *natural (sufficient) statistics*, given the observations $\mathbf{X}$ and the current parameters θ, η.
- *M-step:* Given the expectation of the natural statistics, update the parameters θ, η so as to increase the conditional expectation of the modified log-posterior of the so-called *complete data* [20]. This step is implemented via a loop over all $J \times 9$ parameter subsets $\theta_{j,k}$ specified in Table II. Each subset, depending whether it is adaptive (partially adaptive) or fixed, is updated (partially updated) or not in turn using suitable update rules inspired by [9], [13], [14].

1) Preliminaries:

a) Additive noise and simulated annealing: As explained in [13], where a similar GEM algorithm is used, the mixing parameters $\mathbf{A}_{fn}$ (see Eq. (7)) updated via this GEM algorithm can become stuck into a suboptimal value. To overcome this issue, we use a form of *simulated annealing* proposed in [13], which consists in adding to (7) a noise term whose variance is decreased by a fixed amount at each iteration. Thus, we assume that there is a $J+1$-th source with full-rank time-invariant spatial covariance $\mathbf{\Sigma}_{\mathbf{b},fn} = \sigma_f^2\,\mathbf{I}_I = \mathbf{R}_{J+1,fn}$ and trivial spectral power ($v_{J+1,fn} = 1$) that represents a controllable additive isotropic noise $\mathbf{b}_{fn} = \mathbf{y}_{J+1,fn}$. Introducing this noise component leads to considering the noise covariance $\mathbf{\Sigma}_{\mathbf{b},fn}$ as part of the model parameters θ and to adding it to the mixing equation (7):

$$\mathbf{x}_{fn} = \mathbf{A}_{fn}\mathbf{s}_{fn} + \mathbf{b}_{fn}. \quad (17)$$

b) Complete data log-posterior and natural statistics: We chose $\mathbf{Z} = \{\mathbf{X},\mathbf{S}\}$ as the complete data, where $\mathbf{S} = \{\mathbf{s}_{fn}\}_{f,n}$, and the modified log-posterior of the complete data can be written as:

$$\begin{aligned}\widehat{\mathcal{L}}(\theta,\eta|\mathbf{X},\mathbf{S}) &\stackrel{\text{c}}{=} \widehat{\mathcal{L}}(\mathbf{X}|\mathbf{S};\theta) + \widehat{\mathcal{L}}(\mathbf{S}|\theta) + \log p(\theta|\eta)\\ &\stackrel{\text{c}}{=} -\sum_{f,n}\text{tr}\left[\mathbf{\Sigma}_{\mathbf{b},fn}^{-1}\left(\mathbf{R}_{\mathbf{x},fn} - \mathbf{A}_{fn}\mathbf{R}_{\mathbf{xs},fn}^H - \mathbf{R}_{\mathbf{xs},fn}\mathbf{A}_{fn}^H + \mathbf{A}_{fn}\mathbf{R}_{\mathbf{s},fn}\mathbf{A}_{fn}^H\right)\right] - \sum_{f,n}\log|\mathbf{\Sigma}_{\mathbf{b},fn}|\\ &\quad - \sum_j R_j \sum_{f,n} d_{IS}(\xi_{j,fn}|v_{j,fn}) + \sum_{j,k=1}^{J,9}\log p(\theta_{j,k}|\eta_{j,k}), \quad (18)\end{aligned}$$

where $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$ is the Itakura-Saito (IS) divergence [9], $v_{j,fn}$ are the entries of matrix $\mathbf{V}_j$ specified by (13), and $\mathbf{R}_{\mathbf{x},fn}$, $\mathbf{R}_{\mathbf{xs},fn}$, $\mathbf{R}_{\mathbf{s},fn}$ and $\xi_{j,fn}$ are defined as:

$$\mathbf{R}_{\mathbf{x},fn} \triangleq \widehat{\mathbf{R}}_{\mathbf{x},fn} = \widehat{\mathbb{E}}[\mathbf{x}_{fn}\mathbf{x}_{fn}^H], \quad \mathbf{R}_{\mathbf{xs},fn} \triangleq \widehat{\mathbb{E}}[\mathbf{x}_{fn}\mathbf{s}_{fn}^H], \quad (19)$$

$$\mathbf{R}_{\mathbf{s},fn} \triangleq \widehat{\mathbb{E}}[\mathbf{s}_{fn}\mathbf{s}_{fn}^H], \quad \xi_{j,fn} \triangleq \frac{1}{R_j}\sum_{r=1}^{R_j}\widehat{\mathbb{E}}[|s_{jr,fn}|^2]. \quad (20)$$

It can be easily shown from (18) that the family of functions $\{\exp\widehat{\mathcal{L}}(\mathbf{X},\mathbf{S}|\theta)\}_\theta$ forms an *exponential family* [7], [20], and the set $\mathbb{T}(\mathbf{X},\mathbf{S}) = \{\mathbf{R}_{\mathbf{x},fn},\mathbf{R}_{\mathbf{xs},fn},\mathbf{R}_{\mathbf{s},fn}\}_{f,n}$ is a *natural (sufficient) statistics* [7] for this family. Given this result, we derive a GEM algorithm that is summarized below.

2) Conditional expectation of the natural statistics (E-step): The conditional expectations of the natural statistics $\mathbb{T}(\mathbf{X},\mathbf{S})$ are computed as follows:

$$\widehat{\mathbf{R}}_{\mathbf{xs},fn} = \widehat{\mathbf{R}}_{\mathbf{x},fn}\mathbf{\Omega}_{\mathbf{s},fn}^H, \quad (21)$$

$$\widehat{\mathbf{R}}_{\mathbf{s},fn} = \mathbf{\Omega}_{\mathbf{s},fn}\widehat{\mathbf{R}}_{\mathbf{x},fn}\mathbf{\Omega}_{\mathbf{s},fn}^H + (\mathbf{I}_R - \mathbf{\Omega}_{\mathbf{s},fn}\mathbf{A}_{fn})\mathbf{\Sigma}_{\mathbf{s},fn}, \quad (22)$$

where

$$\mathbf{\Omega}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_{fn}^H\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}, \quad (23)$$

$$\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_{fn}\mathbf{\Sigma}_{\mathbf{s},fn}\mathbf{A}_{fn}^H + \mathbf{\Sigma}_{\mathbf{b},fn}, \quad (24)$$

$$\mathbf{\Sigma}_{\mathbf{s},fn} = \text{diag}\left([\phi_{r,fn}]_{r=1}^R\right), \quad (25)$$

and $\phi_{r,fn} = v_{j,fn}$ if and only if $r \in \mathcal{R}_j$, where $\mathcal{R}_j$ denotes the set of sub-source indices associated with source j in the vector $\mathbf{s}_{fn}$ (see section III-D).

3) Update of the spatial covariances (M-step):

a) Unconstrained time-invariant mixing parameters: We first consider the case where there are no probabilistic priors specified for the mixing parameters $\{\mathbf{A}_j\}_j$ and these parameters are time-invariant. Let $\mathcal{A} \subset \{1,\ldots,R\}$ be a subset of indices of size $\mathcal{D} = \#(\mathcal{A})$. Below we denote by $\mathbf{A}_{fn}^{\mathcal{A}}$, $\widehat{\mathbf{R}}_{\mathbf{xs},fn}^{\mathcal{A}}$ and $\widehat{\mathbf{R}}_{\mathbf{s},fn}^{\mathcal{A}}$ the matrices of respective sizes $I \times \mathcal{D}$, $I \times \mathcal{D}$ and $\mathcal{D} \times \mathcal{D}$, that consist of the corresponding entries of the matrices $\mathbf{A}_{fn}$, $\widehat{\mathbf{R}}_{\mathbf{xs},fn}$ and $\widehat{\mathbf{R}}_{\mathbf{s},fn}$, i.e., $\mathbf{A}_{fn}^{\mathcal{A}} = [\mathbf{A}_{fn}(i,r)]_{i=1,r\in\mathcal{A}}^I$, $\widehat{\mathbf{R}}_{\mathbf{xs},fn}^{\mathcal{A}} = [\widehat{\mathbf{R}}_{\mathbf{xs},fn}(i,r)]_{i=1,r\in\mathcal{A}}^I$, and $\widehat{\mathbf{R}}_{\mathbf{s},fn}^{\mathcal{AB}} = [\widehat{\mathbf{R}}_{\mathbf{s},fn}(r,r')]_{r\in\mathcal{A},r'\in\mathcal{B}}$. We also denote by $\overline{\mathcal{A}} = \{1,\ldots,R\}\backslash\mathcal{A}$ the complementary set. Let $\mathcal{C} \subset \{1,\ldots,R\}$ (resp. $\mathcal{I} \subset \{1,\ldots,R\}$) be the indices of convolutively (resp. instantaneously) mixed sources with adaptive mixing parameters. With these conventions the mixing parameters are updated

as follows [6]:

$$\mathbf{A}_{fn}^{\mathcal{C}} = \left[\sum_{\tilde{n}} \left\{\widehat{\mathbf{R}}_{\mathbf{xs},f\tilde{n}}^{\mathcal{C}} - \mathbf{A}_{f\tilde{n}}^{\overline{\mathcal{C}}} \widehat{\mathbf{R}}_{\mathbf{s},f\tilde{n}}^{\overline{\mathcal{C}}\mathcal{C}}\right\}\right] \left[\sum_{\tilde{n}} \widehat{\mathbf{R}}_{\mathbf{s},f\tilde{n}}^{\mathcal{C}\mathcal{C}}\right]^{-1}, \tag{26}$$

$$\mathbf{A}_{fn}^{\mathcal{I}} = \Re\left[\sum_{\tilde{f},\tilde{n}} \left\{\widehat{\mathbf{R}}_{\mathbf{xs},\tilde{f}\tilde{n}}^{\mathcal{I}} - \mathbf{A}_{\tilde{f}\tilde{n}}^{\overline{\mathcal{I}}} \widehat{\mathbf{R}}_{\mathbf{s},\tilde{f}\tilde{n}}^{\overline{\mathcal{I}}\mathcal{I}}\right\}\right] \left[\Re\left\{\sum_{\tilde{f},\tilde{n}} \widehat{\mathbf{R}}_{\mathbf{s},\tilde{f}\tilde{n}}^{\mathcal{I}\mathcal{I}}\right\}\right]^{-1} \tag{27}$$

b) Other constraints: Estimating time-varying mixing parameters without any priors does not make much sense in practice due to highly unconstrained nature of such the estimation. If the mixing parameters are given some Gaussian priors, closed-form updates similar to (26), (27) can be still derived, since the modified log-posterior (18) will be a quadratic form with respect to the mixing parameters. In case of nongaussian priors some Newton-like updates [22] can be derived.

4) Update of the spectral power parameters (M-step):

a) Unconstrained nonnegative matrices: Let $\mathbf{C}_j = \theta_{j,k}$ ($k = 2, \ldots, 9$) an adaptive or partially adaptive nonnegative matrix (see Tab II) with a uniform prior $p(\theta_{j,k}|\eta_{j,k}) = 1$. Whatever the matrix $\mathbf{C}_j$, it can be shown that the decomposition (13) can be rewritten as $\mathbf{V}_j = (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)\odot\mathbf{E}_j$, where $\mathbf{B}_j$, $\mathbf{D}_j$ and $\mathbf{E}_j$ are some nonnegative matrices that are assumed to be fixed while $\mathbf{C}_j$ is updated. For example, if $\mathbf{C}_j = \mathbf{H}_j^{\text{ft}}$ in (13), one can choose $\mathbf{B}_j = \mathbf{W}_j^{\text{ft}}\mathbf{U}_j^{\text{ft}}\mathbf{G}_j^{\text{ft}}$, $\mathbf{D}_j = \mathbf{I}_N$ and $\mathbf{E}_j = \mathbf{W}_j^{\text{ex}}\mathbf{U}_j^{\text{ex}}\mathbf{G}_j^{\text{ex}}\mathbf{H}_j^{\text{ex}}$. With these notations it can be shown that the conditional expectation of the modified log-posterior (18) of the complete data is non-decreasing when the corresponding update for $\mathbf{C}_j$ does not increase the following cost function:

$$\mathcal{D}_{IS}(\mathbf{C}_j) = \sum\nolimits_{f,n} d_{IS}([\widehat{\mathbf{\Xi}}_j]_{f,n}|[\mathbf{V}_j]_{f,n}), \tag{28}$$

where $\mathbf{V}_j = (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)\odot\mathbf{E}_j$ and $\widehat{\mathbf{\Xi}}_j = [\hat{\xi}_{j,fn}]_{f,n}$ with $\hat{\xi}_{j,fn}$ computed as follows:

$$\hat{\xi}_{j,fn} = \frac{1}{R_j}\sum\nolimits_{r\in\mathcal{R}_j} \widehat{\mathbf{R}}_{\mathbf{s},fn}(r,r), \tag{29}$$

where $\widehat{\mathbf{R}}_{\mathbf{s},fn}$ is computed in (22) and $\mathcal{R}_j$ is defined at the end of Section IV-B2. Applying some standard derivations (see, e.g., [9]), one can obtain the following nonnegative MU rule [7]

$$\mathbf{C}_j = \mathbf{C}_j \odot \frac{\mathbf{B}_j^T[\widehat{\mathbf{\Xi}}_j \odot \mathbf{E}_j \odot \{(\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)\odot\mathbf{E}_j\}^{\cdot -2}]\mathbf{D}_j^T}{\mathbf{B}_j^T[\mathbf{E}_j \odot \{(\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)\odot\mathbf{E}_j\}^{\cdot -1}]\mathbf{D}_j^T} \tag{30}$$

that guarantees non-increase of the cost function (28), and thus non-decrease of the conditional expectation of the modified log-posterior (18) of the complete data. These update rules, as applied to multichannel audio, are in fact a generalization of the GEM-MU algorithm proposed in [21], that has been shown to converge much more quickly than the GEM algorithm in [13].

b) Discrete state-based constraints: Let us now assume that $\theta_{j,4} = \mathbf{G}_j^{\text{ex}}$ is subject to a discrete state-based constraint (similarly for $\theta_{j,8} = \mathbf{G}_j^{\text{ft}}$). Note that when time-localized patterns $\mathbf{H}_j^{\text{ex}}$ (or $\mathbf{H}_j^{\text{ft}}$) have non-zero overlaps in time of maximum length L (see, e.g., Fig. 4) the model becomes equivalent to an HMM of the order L (in case of GMMs) or of the order $L+1$ (in case of HMMs). In order to avoid the complications of requiring consistency of overlapping patterns (which would introduce temporal constraints somewhat reminiscent of an HMM), in our baseline implementation and in the updates described below we only consider non-overlapping time-localized patterns $\mathbf{H}_j^{\text{ex}} = \mathbf{I}_N$ in case of discrete state-based constraints. The updates are performed as follows:

1) Set $\tilde{\mathbf{G}}_j^{\text{ex}} = \mathbf{G}_j^{\text{ex}}$, and fill each entry of each column of $\tilde{\mathbf{G}}_j^{\text{ex}}$ with the nonzero entry of the respective column of $\mathbf{G}_j^{\text{ex}}$.
2) If $\mathbf{G}_j^{\text{ex}}$ is adaptive, do for every $k = 1, \ldots, K_j^{\text{ex}}$:
   - Set $\mathbf{C}_j = \tilde{\mathbf{G}}_j^{\text{ex}}$, and set all the elements of $\mathbf{C}_j$ to zero, except the k-th row.
   - Update $\mathbf{C}_j$ using several iterations of (30) [8].
   - Set the k-th row of $\tilde{\mathbf{G}}_j^{\text{ex}}$ equal to that of $\mathbf{C}_j$.
3) For every $k = 1, \ldots, K_j^{\text{ex}}$ and $m = 1, \ldots, M_j^{\text{ex}}$ set $\mathbf{C}_j = \tilde{\mathbf{G}}_j^{\text{ex}}$, set all the elements of $\mathbf{C}_j$ to zero, except the (k,m)-th one, and compute the IS divergence $\mathcal{D}_{IS}(k,m)$ between $\mathbf{V}_j = (\mathbf{B}_j\mathbf{C}_j\mathbf{D}_j)\odot\mathbf{E}_j$ and $\widehat{\mathbf{\Xi}}_j$, as in (28).
4) Update the state sequence $\mathbf{q}_j^{\text{ex}}$ using the Viterbi algorithm [45] to minimize the following criterion:

$$\mathbf{q}_j^{\text{ex}} = \arg\min_{\mathbf{q}_j^{\text{ex}}} \sum_{m=2}^{M_j^{\text{ex}}} \mathcal{D}_{IS}(q_{j,m}^{\text{ex}}, m) - \log p(\mathbf{q}_j^{\text{ex}}|\mathbf{\Lambda}_j^{\text{ex}}),$$

   where $p(\mathbf{q}_j^{\text{ex}}|\mathbf{\Lambda}_j^{\text{ex}})$ is computed as in (14).
5) Set $\mathbf{G}_j^{\text{ex}} = \tilde{\mathbf{G}}_j^{\text{ex}}$ and set to zero all the entries of $\mathbf{G}_j^{\text{ex}}$, except those corresponding to $\mathbf{q}_j^{\text{ex}}$.
6) If $\mathbf{\Lambda}_j^{\text{ex}}$ is adaptive, update the transition probabilities as $\lambda_{j,kk'}^{\text{ex}} = \frac{\sum_{m=2}^{M_j^{\text{ex}}} \mathbf{1}(q_{j,m-1}^{\text{ex}}=k, q_{j,m}^{\text{ex}}=k')}{(M_j^{\text{ex}}-1)\sum_{m=2}^{M_j^{\text{ex}}} \mathbf{1}(q_{j,m-1}^{\text{ex}}=k)}$ in case of HMM or S-HMM or as $\lambda_{j,kk'}^{\text{ex}} = \frac{1}{M_j^{\text{ex}}-1}\sum_{m=2}^{M_j^{\text{ex}}} \mathbf{1}(q_{j,m}^{\text{ex}} = k')$ in case of GMM or GSMM.

c) Other constraints: We here discuss the updates that are not yet included in our current baseline implementation (see Sec. II-D).

An EM algorithm update rules for time pattern weights $\mathbf{G}_j^{\text{ex}}$ or $\mathbf{G}_j^{\text{ft}}$ with time continuity priors, such as inverse-Gamma or Gamma Markov chain priors, can be found in [9]. However, one cannot use these rules within our GEM algorithm, since we use a different, reduced, complete data set, as compared

[6] We see that the mixing parameters for different sources are updated jointly by Eqs. (26), (27), while we have claimed in the beginning of Section IV that they will be updated in an alternated manner. However, since we can here update parameters jointly without loss of flexibility, we do so, since joint optimization, as compared to the alternated one, leads in general to a faster convergence.

[7] In the case of partially adaptive matrix $\mathbf{C}_j$, only the adaptive matrix entries are updated with rule (30).

[8] Several iterations of update rule (30) are needed because all entries of $\tilde{\mathbf{G}}_j^{\text{ex}}$ are initialized in step 1 from a particular sequence of gains carried by $\mathbf{G}_j^{\text{ex}}$ and optimized for the current state sequence $\mathbf{q}_j^{\text{ex}}$. Performing only one update of (30) would unfavor state sequence evaluation. However, to avoid all these issues, in our implementation we just keep matrix $\tilde{\mathbf{G}}_j^{\text{ex}}$ in memory, skip step 1, and do only one iteration of (30).

to the one used in [9]. Nevertheless, one can always use some Newton-like updates [22] for these priors.

If a matrix $\theta_{j,k}$ ($k = 2, \ldots, 9$) is constrained with a sparsity-inducing prior [4], such as a Laplacian prior (corresponding to an l_1 norm penalty), it can be updated using the multiplicative updates described in [46], [47]. However, in such a case the renormalization described in the subsection below could not be applied, since it would change the value of the optimized criterion (16). At the same time, without any renormalization, the sparsity-inducing prior would loose its influence. To avoid that, all the other parameter subsets $\theta_{j,l}$ ($l \neq k$) should be constrained, e.g., to have a unitary (say l_1) norm, which can be handled using the gradient descent updates from [46] or the modified multiplicative updates from [47].

5) Renormalization: At the end of each GEM iteration, in order to avoid numerical (under/over-flow) problems, a renormalization of some parameters is done if needed, i.e., if these parameters are not already constrained by some priors that are not scale-invariant. This procedure is similar to the one described in [13], and it does not change the value of the optimized criterion (16). For example, the columns of matrix $\mathbf{U}_j^{\text{ex}}$ can be divided by their energies, and the rows of $\mathbf{G}_j^{\text{ex}}$ scaled accordingly (see (13)). Similar renormalization is applied in turn to each patameter subsets pairs $\theta_{j,k}, \theta_{j,k+1}$ ($k = 1, \ldots, 8$), and at the end of this operation the total energy is relegated into $\theta_{j,9}$.

C. Source estimation

Given the estimated model parameters θ, the sources can be estimated in the minimum mean square error (MMSE) sense via the Wiener filtering:

$$\hat{\mathbf{y}}_{j,fn} = v_{j,fn}\mathbf{R}_{j,fn}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{x}_{fn}, \tag{31}$$

where $\boldsymbol{\Sigma}_{\mathbf{x},fn} = \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$. The counterpart of this equation for quadratic TF representations is given in Appendix A.

V. Experimental Illustrations

The goals of this experimental part are to illustrate on some examples how to specify the prior information in the framework, given a particular source separation problem, and to demonstrate that we can implement the existing and new methods within the framework. For that we first give an example of application of the framework to a music recording in a *non-blind* setting, i.e., when different sources are given different models according to the prior information. Second, we consider a few blind framework instances, corresponding to existing and new methods, and apply them for separation of underdetermined speech and music mixtures. Third, we describe how to apply the framework to solve the source separation problem mentioned in the beginning of the introduction, i.e., the separation of bass, drums and melody in music recordings. Finally, we briefly mention our application of the framework for speech separation in the context of noise robust speech recognition.

A. Non-blind separation of one music recording

1) Data: As an example stereo music recording to separate we took the 23-second snip of the song "Que pena tanto faz" by Tamy from the test dataset of the SiSEC 2008 [30] "Professionally produced music recordings" task. We know about this recording that there are two sources, a female singing voice and a guitar, that the voice is instantaneously mixed (panned) in the middle [9] and the guitar is possibly a non-point convolutive source.

2) Constraint specification and parameter initialization: To account for this information within our framework, we have chosen the following constraints. The singing voice mixing parameters $\mathbf{A}_1$ form a fixed tensor of size $2 \times 1 \times F \times N$ with all entries equal to 1. The guitar mixing parameters $\mathbf{A}_2$ form an adaptive tensor of size $2 \times 2 \times F \times N$ subject to the time-invariance constraint. The spectral powers $\mathbf{V}_j$ ($j = 1, 2$) are constrained to $\mathbf{V}_j = \mathbf{W}_j^{\text{ex}}\mathbf{U}_j^{\text{ex}}\mathbf{G}_j^{\text{ex}}\mathbf{H}_j^{\text{ex}}$ [5] with $\mathbf{W}_j^{\text{ex}}$ and $\mathbf{H}_j^{\text{ex}}$ being fixed, and $\mathbf{U}_j^{\text{ex}}$ and $\mathbf{G}_j^{\text{ex}}$ being adaptive. The narrowband spectral patterns $\mathbf{W}_j^{\text{ex}}$ include $6 \times L$ harmonic patterns modeling the harmonic part of L pitches and 9 smooth patterns (see Fig. 4 (E) and [14]). The L pitches are chosen to cover the range of 77 - 1397 Hz (39 - 89 on the MIDI scale), which is enough for both the guitar and this particular singing. The time-localized patterns $\mathbf{H}_1^{\text{ex}}$ and $\mathbf{H}_2^{\text{ex}}$ are different. The singing voice time-localized patterns $\mathbf{H}_1^{\text{ex}}$ include half-Gaussians truncated at the left, i.e., only the right half is kept. The guitar time-localized patterns $\mathbf{H}_2^{\text{ex}}$ include decreasing exponentials to model the decay part of the notes and discrete Dirac functions to model note attacks (see Fig. 4 (H)). All adaptive parameters are initialized with random values. Finally, we used the ERB quadratic representation described in [18] as signal representation.

3) Results: After 500 iterations of the proposed GEM algorithm the separation results, measured in terms of the source to distortion ratio (SDR) [48], were 7.2 and 8.9 dB for voice and guitar, respectively. We have also separated the same mixture using all the blind settings described in the following section. The best results of 5.5 and 7.1 dB SDR were obtained by the unconstrained NMF spectral power model with the instantaneous rank-1 mixing, i.e., by the multichannel NMF for instantaneous mixtures [13].

4) Discussion: We see that our informed setting outperforms any blind setting by at least 1.7 dB SDR. This improvement is essentially due to the combination of rank-1 instantaneous and full-rank convolutive mixing models and the information about the position of one source. Moreover, while it is common in professionally produced music recordings that some sources are mixed instantaneously (panned) and others convolutively (e.g., live-recorded tracks or some artificial reverberation is added), in our best knowledge such hybrid models were not yet proposed for audio source separation, and it now becomes possible to implement them within our framework.

[9] This information can be for example obtained by subtracting the left channel from the right one and checking that the voice is cancelled.

B. Blind separation of underdetermined speech and music mixtures

1) Data: Here we evaluate several settings of our framework on the development dataset of the SiSEC 2010 [29] "Underdetermined-speech and music mixtures" task. This dataset include 10-seconds length instantaneous, convolutive and live-recorded stereo mixtures of three or four music and speech sources (see [29] for more details).

2) Constraint specification and parameter initialization: We consider eight blind settings of the framework that are specified by the following constraints. For all settings and for all sources $\mathbf{A}_j$ forms an adaptive tensor of size $2 \times R_j \times F \times N$ subject to the time-invariance constraint and subject to the frequency invariance constraint for instantaneous mixtures only. The spectral power of each source is structured as $\mathbf{V}_j = \mathbf{E}_j^{\mathrm{ex}} \mathbf{P}_j^{\mathrm{ex}}$ [5]. The eight settings are generated by all possible combinations of the following possibilities (see also Table IV):

- *Rank:* The rank R_j is either 1 or 2 (full-rank).
- *Spectral structure:* The characteristic spectral patterns $\mathbf{E}_j^{\mathrm{ex}}$ are either *unconstrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{W}_j^{\mathrm{ex}}$ with adaptive $\mathbf{W}_j^{\mathrm{ex}}$, or *constrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \bar{\mathbf{W}}_j^{\mathrm{ex}} \mathbf{U}_j^{\mathrm{ex}}$ with fixed $\bar{\mathbf{W}}_j^{\mathrm{ex}}$ being composed of harmonic and noise-like and smooth narrowband spectral patterns (see Fig. 4 (E) and [14]), and adaptive $\mathbf{U}_j^{\mathrm{ex}}$ (see Fig. 4 (F)) that is very sparse so as to eliminate invalid combinations of narrowband spectral patterns (e.g., patterns corresponding to different pitches should not be combined together).
- *Temporal structure:* The time activation coefficients $\mathbf{P}_j^{\mathrm{ex}}$ are either *unconstrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{G}_j^{\mathrm{ex}}$ with adaptive $\mathbf{G}_j^{\mathrm{ex}}$, or *constrained*, i.e., $\mathbf{E}_j^{\mathrm{ex}} = \mathbf{G}_j^{\mathrm{ex}} \bar{\mathbf{H}}_j^{\mathrm{ex}}$ with fixed $\bar{\mathbf{H}}_j^{\mathrm{ex}}$ being composed of decreasing exponentials, as those on Fig. 4 (H), and adaptive $\mathbf{G}_j^{\mathrm{ex}}$.

The two settings with $R_j = 1$ and 2, and unconstrained $\mathbf{E}_j^{\mathrm{ex}}$ and $\mathbf{P}_j^{\mathrm{ex}}$ correspond to the state-of-the-art methods [13] and [17], respectively (see Section III-F), while the remaining six settings are new.

In line with [13], parameter estimation via GEM is sensitive to initialization for all the settings we consider. To provide our GEM algorithm with a "good initialization" we used for the instantaneous mixtures the DEMIX mixing matrix estimation algorithm [49] to initialize mixing parameters $\mathbf{A}_j$, followed by l_0 norm minimization (see e.g., [1]) and Kullback-Leibler (KL) divergence minimization (see [13]) to initialize the source power spectra $\mathbf{V}_j$. For synthetic convolutive and live recorded mixtures we first estimated the time differences of arrival (TDOAs) using the MVDRW estimation algorithm proposed in [50], that is based on a variance distortionless response (MVDR) beamformer. The estimated TDOAs were then used to initialize anechoic mixing parameters $\mathbf{A}_j$, followed by binary masking and KL divergence minimization (see [13]) to initialize the source power spectra $\mathbf{V}_j$. As signal representation we used the STFT.

3) Results: Source separation results in terms of average SDR after 200 iterations of the proposed GEM algorithm are summarized in Table IV together with results of the *baseline* used for initialization.

4) Discussion: As expected, in most cases rank-1 spatial covariances perform the best for instantaneous mixtures and full-rank spatial covariances perform the best for synthetic convolutive and live recorded mixtures. Moreover, in all the cases there is at least one of the six new methods that outperforms the state-of-the-art methods [13] and [17]. One can note that for music sources constraining the spectral structure does not improve the separation performance [10], however, constraining the temporal structure does improve it. For speech sources constraining both the spectral and the temporal structures improves the separation performance in most cases. This is probably because the unconstrained NMF is a poor model for speech. Indeed, as compared to simple music, speech includes much more different spectral patterns, notably due to a more pronounced vibrato effect (varying pitch). As a consequence, the unconstrained NMF model needs much more components to describe this variability, thus it cannot be estimated in a robust way from these quite short 10-second length mixtures. Introducing spectral and temporal constraints makes model estimation more robust.

C. Separation of bass, drums and melody in music recordings

Here we describe how to apply our framework to the separation of the bass, the drums, the melody and the remaining instruments from a stereo professionally produced music recording. This source separation problem is of great practical interest for music information retrieval and remastering (e.g., karaoke) applications.

1) State-of-the-art: The state-of-the-art approaches targeting this problem suffer from the following limitations. First, existing drum [52] and melody [16] separation algorithms have been designed for single-channel (mono) recordings and may fail to segregate the melody from the other harmonic sources despite the fact that they have different spatial directions. Second, blind source separation methods relying on joint use of spatial and spectral diversity, such as, e.g., the multichannel NMF [13], need some user input to label separated signals [21] and cannot separate sources mixed in the same direction, which is a very common situation, e.g., for singing melody and drums. Finally, no state-of-the-art approach treats this problem in a joint fashion and cascading the methods (e.g., separating the drums, then separating the melody, etc.) is clearly suboptimal. Thus, it is clear that an efficient solution to this problem should rely on:

- some prior knowledge about the source spectral characteristics (to label the sources automatically),
- the spatial diversity of different sources,
- some model describing harmonicity, and
- joint modeling of all sources.

2) Constraint specification, parameter initialization and reconstruction: Our framework satisfies these requirements, and in order to account for this information we have chosen the following constraints. The two-channel mixture is modeled as a sum of 12 sources: 4 sources ($j = 1, \ldots, 4$) representing

[10]The results for synthetic convolutive mixtures of music sources are not very informative because of the poor overall performance.

Mixing				instantaneous		synthetic convolutive				live recorded			
Sources				speech	music	speech		music		speech		music	
Microphone spacing				-	-	5 cm	1 m	5 cm	1 m	5 cm	1 m	5 cm	1 m
Number of 10 second-length mixtures				6	4	10	10	4	4	10	10	4	4
baseline (l_0 minimization [51] or binary masking)				8.6	12.4	1.0	1.4	-0.9	-0.7	1.1	1.4	2.5	0.3
Method	rank R_j	spectral struct.	temporal struct.										
[13]	1	unconstrained	unconstrained	8.8	17.2	1.6	2.1	-1.1	-1.2	2.2	2.5	3.2	0.4
[17]	2	unconstrained	unconstrained	8.9	17.0	1.8	2.7	-0.5	-0.2	2.0	3.0	3.5	0.8
new	1	constrained	unconstrained	**10.5**	13.6	1.9	2.5	-0.5	-0.5	2.2	2.8	3.0	0.5
new	2	constrained	unconstrained	10.4	13.0	**2.1**	3.1	-0.7	-0.4	2.3	3.2	3.2	0.8
new	1	unconstrained	constrained	8.9	**18.6**	1.5	2.2	-0.8	-0.5	2.4	2.6	3.4	0.9
new	2	unconstrained	constrained	8.7	15.4	1.8	2.6	-0.4	0.0	2.1	2.9	**4.5**	**1.8**
new	1	constrained	constrained	**10.5**	15.7	**2.1**	2.9	-1.2	**0.3**	**2.5**	3.9	3.2	0.4
new	2	constrained	constrained	10.2	13.8	**2.1**	**4.5**	**0.0**	-0.3	2.3	**5.0**	3.7	1.0

TABLE IV
AVERAGE SDRS ON SUBSETS OF SISEC 2010 "UNDERDETERMINED SPEECH AND MUSIC MIXTURES" TASK DEVELOPMENT DATASET.

the bass, 4 sources ($j = 5, \ldots, 8$) representing the drums [11], and the remaining 4 sources ($j = 9, \ldots, 12$) representing the melody and the other instruments. Each set of mixing parameters $\mathbf{A}_j$ ($j = 1, \ldots, 12$) form an adaptive tensor of size $2 \times 2 \times F \times N$ subject to the time-invariance constraint. The spectral powers $\mathbf{V}_j$ of the bass and the drums ($j = 1, \ldots, 8$) are constrained to $\mathbf{V}_j = \mathbf{W}_j^{\mathrm{ex}} \mathbf{G}_j^{\mathrm{ex}\ 5}$ with $\mathbf{G}_j^{\mathrm{ex}}$ being adaptive and $\mathbf{W}_j^{\mathrm{ex}}$ being fixed and pre-trained (using our framework) from isolated bass and drum samples from the RWC music database [53]. The spectral powers $\mathbf{V}_j$ of the melody and the remaining instruments ($j = 9, \ldots, 12$) are constrained to $\mathbf{V}_j = \mathbf{W}_j^{\mathrm{ex}} \mathbf{U}_j^{\mathrm{ex}} \mathbf{G}_j^{\mathrm{ex}\ 5}$ with $\mathbf{W}_j^{\mathrm{ex}}$ being fixed, and $\mathbf{U}_j^{\mathrm{ex}}$ and $\mathbf{G}_j^{\mathrm{ex}}$ being adaptive. The narrowband spectral patterns $\mathbf{W}_j^{\mathrm{ex}}$ ($j = 9, \ldots, 12$) include $3 \times L$ harmonic patterns modeling the harmonic part of L pitches (see [14]). The L pitches are chosen to cover the range of 27 - 4186 Hz (21 - 108 on the MIDI scale), which is enough to cover the pitch range of most instruments. All adaptive parameters are initialized with random values, except the mixing parameters $\mathbf{A}_j$ ($2 \times 2 \times F \times N$ tensors) that are initialized with the same (random) $2 \times 2 \times N$ tensor for all frequency bins. We used the ERB quadratic representation in [18] as signal representation due to its higher low-frequency resolution than the STFT, which is desirable for the modeling of bass sounds. Once the GEM algorithm has run, the 12 sources are estimated via Wiener filtering. The bass and the drums are reconstructed by summing the corresponding source estimates, the melody is reconstructed by choosing the most energetic source among the corresponding four ($j = 9, \ldots, 12$) sources, and the remaining instruments by summing the other three sources.

3) Results: The corresponding source separation script together with one separation example are available from the FASST web page [25]. Note that this example is a difficult, real-world mixture, which involves several sources mixed in the center (bass, singing voice, certain drums) and several harmonic sources with comparable pitch range (singing voice, piano).

[11]The bass is modeled as a sum of 4 sources to facilitate initialization, since we do not know a priori its spatial direction. The drums are modeled as a sum of 4 sources for the same reason, but also because the drum track is often composed of several sources (e.g., snare, hi-hat, cymbals, etc) that can be mixed in different directions.

D. Separation of speech in multi-source environment for noise robust speech recognition

We have also applied the framework for the problem of speech separation in reverberant noisy multi-source environment. This was done for our submission to the 2011 CHiME Speech Separation and Recognition Challenge [12]. The corresponding description can be found in [54] and some separation examples are available from a demo web page at [13].

VI. CONCLUSION

We have introduced a general flexible audio source separation framework that generalizes several existing source separation methods, brings them into a common framework, and allows to imagine and implement new efficient methods, given the prior information about a particular source separation problem. Besides the framework itself, we proposed a new temporal structure for NMF-like decompositions and an original mixing model formulation combining rank-1 and full-rank spatial mixing models in a homogeneous way. Finally, we provided a proper probabilistic formulation of local Gaussian modeling for quadratic time-frequency representations.

In the experimental part we have illustrated how to specify the prior information about a particular source separation problem within the framework, and we have shown that the framework allows implementing existing and new efficient source separation methods. We have also demonstrated that in some situations our new propositions can improve the source separation performance, as compared to the state-of-the-art. As such combining instantaneous rank-1 and and convolutive full-rank can be useful for separation of professionally produced music recordings, and the newly proposed temporal structure for NMF-like decompositions brings some improvement for blind separation of underdetermined mixtures of speech and music sources.

As for further research, the following extensions could be introduced to the framework. In a similar fashion as for

[12]http://spandh.dcs.shef.ac.uk/projects/chime/challenge.html
[13]http://www.irisa.fr/metiss/ozerov/chime_ssep_demo.html

spectral power, a flexible structure can be specified for the mixing parameters. E.g., the time-varying mixing parameters could be represented in terms of time-localized and locally time-invariant mixing parameter patterns, thus allowing the modeling of moving sources. Another interesting extension would be to introduce possible coupling between parameter subsets, thus allowing, e.g., the representation of the characteristic spectral patterns of different sources as linear combinations of eigenvoices [55] or eigeninstruments [56]. In fact, some parameter subsets corresponding to different sources can share common properties, and introducing such a coupling would make the estimation of these parameters more robust.

APPENDIX A
PROBABILISTIC FORMULATION OF THE LOCAL GAUSSIAN MODEL FOR QUADRATIC REPRESENTATIONS

Here we give a proper probabilistic formulation of the local Gaussian model (4) for quadratic representations, explaining the exact meaning of the empirical covariance (3) and a justification of the criterion (16).

A. Input representation

Following [10], [18], we assume that the considered quadratic TF representation is computed by local averaging of a linear TF representation such as a STFT or an ERB filterbank. We assume that the indexing of the considered linear TF complex-valued representation, hereafter noted as $m = 1, \ldots, M$, can be in general different from the indexing f, n of the quadratic representation (3). Such a formulation allows considering linear and quadratic representations with different TF resolutions, but also using linear TF representations that do not allow any uniform TF indexing, e.g., an ERB representation with different sampling frequencies in different frequency bands or a signal-adapted multiple-window STFT [57]. The mixing equation (1) now writes as

$$\mathbf{x}_m = \sum\nolimits_{j=1}^{J} \mathbf{y}_{j,m}, \tag{32}$$

and we re-define the empirical covariance (3) as

$$\widehat{\mathbf{R}}_{\mathbf{x},fn} = \sum\nolimits_{m} (\omega_{fn,m}^{\mathrm{ana}})^2 \mathbf{x}_m \mathbf{x}_m^H, \tag{33}$$

where $\omega_{fn,m}^{\mathrm{ana}} \geq 0$, satisfying $\sum_{f,n} (\omega_{fn,m}^{\mathrm{ana}})^2 = 1$, are the coefficients of a local bi-dimensional *analysis* window specifying a neighbourhood of the TF point (f, n) [10], [18].

B. Local Gaussian model

In this setting the local Gaussian model (4) is re-defined as follows. Each vector $\mathbf{y}_{j,m}$ is assumed to be distributed as

$$\mathbf{y}_{j,m} \sim \mathcal{N}_c\left(\bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right) \tag{34}$$

with probability $(\omega_{fn,m}^{\mathrm{ana}})^2$. In other words, $\mathbf{y}_{j,m}$ is a realization of a GMM. Moreover, the vectors $\{\mathbf{y}_{j,m}\}_j$ are assumed to be independent only conditionally on the same GMM state. More precisely, the joint probability density function of $\{\mathbf{y}_{j,m}\}_j$ is defined as

$$p(\mathbf{y}_{1,m}, \ldots, \mathbf{y}_{J,m}) \triangleq \sum\nolimits_{fn} (\omega_{fn,m}^{\mathrm{ana}})^2 \prod\nolimits_j N_c\left(\mathbf{y}_{j,m}; \bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right). \tag{35}$$

C. Model estimation criterion

Under the above-presented assumptions (see (32) and (35)), the log-posterior $\log p(\theta, \eta|\mathbf{X})$, maximized by the MAP criterion, writes

$$\log p(\theta, \eta|\mathbf{X}) \stackrel{\mathrm{c}}{=} \log p(\mathbf{X}|\theta) + \log p(\theta|\eta) = \sum_{f,n} \log \sum_{m} (\omega_{fn,m}^{\mathrm{ana}})^2 N_c(\mathbf{x}_m; \bar{0}, \mathbf{\Sigma}_{\mathbf{x},fn}) + \log p(\theta|\eta), \tag{36}$$

where $\mathbf{\Sigma}_{\mathbf{x},fn} = \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$. Log-posterior (36) is difficult to optimize, due to summations in log-domain. Thus, following the EM methodology [20], we replace $\log p(\theta, \eta|\mathbf{X})$ by its lower bound

$$\sum_{f,n} \sum_{m} (\omega_{fn,m}^{\mathrm{ana}})^2 \log N_c(\mathbf{x}_m; \bar{0}, \mathbf{\Sigma}_{\mathbf{x},fn}) + \log p(\theta|\eta), \tag{37}$$

using Jensen's inequality [20], and we get the criterion (16) with empirical covariances $\widehat{\mathbf{R}}_{\mathbf{x},fn}$ computed as in (33). Thus, the criterion (16) maximizes a lower bound of the log-posterior (36).

Note, that with this formulation we could obtain exactly the same updates as those presented in Section IV-B by deriving a GEM algorithm for the MAP criterion (36). This is because the computing of the lower bound (37) is based on the EM methodology. However, we prefer to keep the criterion (16), since it makes the formulation more compact and links it to quadratic representations and to the existing works [10], [18].

D. Source estimation

The sources can be estimated as follows [10], [18]:

$$\hat{\mathbf{y}}_{j,m} = \sum\nolimits_{f,n} \omega_{fn,m}^{\mathrm{syn}} \omega_{fn,m}^{\mathrm{ana}} v_{j,fn} \mathbf{R}_{j,fn} \mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \mathbf{x}_m, \tag{38}$$

where $\omega_{fn,m}^{\mathrm{syn}} \geq 0$ is a so-called *synthesis* window satisfying $\sum_{f,n} \omega_{fn,m}^{\mathrm{syn}} \omega_{fn,m}^{\mathrm{ana}} = 1$. This estimator becomes the MMSE estimator when $\omega_{fn,m}^{\mathrm{syn}} = \omega_{fn,m}^{\mathrm{ana}}$.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

[1] E. Vincent, M. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010, ch. 7, pp. 162–185.

[2] 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.

[3] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Proceedings of the 7th International Symposium on Signal Processing and its Applications*, 2003, pp. II–73–76.

[4] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by nonnegative sparse coding of power spectra," in *Proc. 5th International Symposium Music Information Retrieval (ISMIR'04)*, Oct. 2004, pp. 318–325.

[5] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '05)*, Mohonk, NY, USA, Oct. 2005, pp. 78–81.

[6] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[7] 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," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[8] R. Blouet, G. Rapaport, I. Cohen, and C. Févotte, "Evaluation of several strategies for single sensor speech/music separation," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Las Vegas, USA, Apr. 2008, pp. 37 – 40.

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

[10] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'09)*, 2009, pp. 775 – 782.

[11] 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.

[12] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, Oct. 18–21, 2009, pp. 121–124.

[13] 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, March 2010.

[14] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[15] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 538–549, 2010.

[16] J. L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[17] 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 *10th Int. Conf. on Information Sciences, Signal Proc. and their applications (ISSPA'10)*, 2010, pp. 1–4.

[18] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Saint-Malo, France, Sep. 27-30 2010, pp. 73–80.

[19] ——, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[21] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, Czech Republic, May 2011, pp. 257 – 260.

[22] J.-F. Cardoso, M. Le Jeune, J. Delabrouille, M. Betoule, and G. Patanchon, "Component separation with flexible models — Application to multichannel astrophysical observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 5, pp. 735–746, 2008.

[23] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience. Hindawi Publishing Corp*, vol. 2008, 2008.

[24] 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.

[25] ——, "Flexible Audio Source Separation Toolbox (FASST)." [Online]. Available: http://bass-db.gforge.inria.fr/fasst/

[26] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Magazine*, vol. 9, no. 2, pp. 21–67, 1992.

[27] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[28] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[29] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong, "The 2010 signal separation evaluation campaign (SiSEC2010): - Audio source separation -," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Saint-Malo, France, Sep. 2010, pp. 114–122.

[30] E. Vincent, S. Araki, and P. Bofilld, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA'09)*, 2009, pp. 734–741.

[31] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97)*, April 1997, pp. 3617 – 3620.

[32] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69 – 84, 2010.

[33] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs." in *Fifth International Conference on Independent Component Analysis*, Granada, Spain, Sep. 2004, pp. 494–499.

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

[35] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'07)*, vol. 1, 2007, pp. 53 – 56.

[36] I. Lee, T. Kim, and T.-W. Lee, "Independent vector analysis for convolutive blind speech separation," in *Blind speech separation*. Springer, 2007, pp. 169–192.

[37] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'08)*, 2008, pp. 1833–1836.

[38] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, no. Article ID 785152, 2009.

[39] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 793–799.

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

[41] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'01)*, San Diego, USA, Dec. 2001, pp. 1–6.

[42] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[43] R. Hennequin, R. Badeau, and B. David, "NMF with time–frequency activations to model nonstationary audio events," *IEEE Transactions on*

Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 744–753, 2011.
[44] Y. Meron and K. Hirose, "Separation of singing and piano sounds," in *Proc. Int. Conf. on Spoken Language Processing*, 1998.
[45] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[46] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
[47] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN'04)*, 2004, pp. 2529–2533.
[48] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[49] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121 –133, jan. 2010.
[50] C. Blandin, E. Vincent, and A. Ozerov, "Multi-source TDOA estimation using SNR-based angular spectra," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, Czech Republic, May 2011, pp. 2616 – 2619.
[51] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'07)*, 2007, pp. 430–437.
[52] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.
[53] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound databases," in *5th International Symposium on Music Information Retrieval (ISMIR)*, 2004, pp. 229–230. [Online]. Available: http://staff.aist.go.jp/m.goto/RWC-MDB/
[54] A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *International Workshop on Machine Listening in Multisource Environments (CHiME 2011)*, Florence, Italy, September 2011, pp. 86–87.
[55] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.
[56] G. Grindlay and D. Ellis, "Multi-voice polyphonic music transcription using eigeninstruments," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, 2009, pp. 53–56.
[57] L. Benaroya, R. Blouet, C. Févotte, and I. Cohen, "Single sensor source separation using multiple-window STFT representation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sep. 12–14 2006.

Alexey Ozerov holds a Ph.D. in Signal Processing from the University of Rennes 1 (France). He worked towards this degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, he received an M.Sc. degree in Mathematics from the Saint-Petersburg State University (Russia) in 1999 and an M.Sc. degree in Applied Mathematics from the University of Bordeaux 1 (France) in 2003. From 1999 to 2002, Alexey worked at Terayon Communicational Systems (USA) as a R&D software engineer, first in Saint-Petersburg and then in Prague (Czech Republic). He was for one year (2007) in Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, and for one year and half (2008-2009) in TELECOM ParisTech / CNRS LTCI - Signal and Image Processing (TSI) Department. Now he is with METISS team of IRISA / INRIA - Rennes as a Post-Doc researcher. His research interests include audio source separation, source coding, and automatic speech recognition.

Emmanuel Vincent (M'07 - SM'10) is a Research Scientist with the French National Institute for Research in Computer Science and Control (INRIA, Rennes, France). Prior to that he received the Ph.D. degree in music signal processing from IRCAM (Paris, France) in 2004 and worked as a Research Assistant with the Centre for Digital Music at Queen Mary, University of London (London, U.K.) from 2004 to 2006. His research focuses on probabilistic machine learning for speech and audio signal processing, with application to real-world audio source localization and separation, noise-robust speech recognition and music information retrieval. He is the founding chair of the annual Signal Separation Evaluation Campaign (SiSEC). Dr. Vincent is an Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing.

Frédéric Bimbot received in 1988 a PhD in signal processing (speech synthesis using temporal decomposition), after graduating as a telecommunication engineer in 1985 (ENST, Paris, France). He also obtained in 1987 a B.A. in Linguistics (Sorbonne Nouvelle University, Paris III).

In 1990, he joined CNRS (French National Center for Scientific Research) as a permanent researcher, worked with ENST for 7 years and then moved to IRISA (CNRS & INRIA), in Rennes. He also repeatedly visited AT&T – Bell Laboratories between 1990 and 1999. He is now a Senior Researcher with CNRS.

He is heading the (IRISA/CNRS & INRIA/Rennes) METISS research group, dedicated to selected topics in speech and audio processing and his research is focused on speech and audio analysis, speaker recognition, music content modeling and audio source separation. He is also in charge of the coordination of D5 Department (Digital Signals and Images, Robotics) at IRISA.

He has been involved in a number of research projects, among which the ongoing QUAERO project in multimedia information retrieval.