Multi-channel audio source separation using multiple deformed references

Nathan Souviraà-Labastie, Anaik Olivero, Emmanuel Vincent, Frédéric Bimbot

Multi-channel audio source separation using multiple deformed references

Nathan Souviraà-Labastie, Anaik Olivero, Emmanuel Vincent, *Senior Member, IEEE,* and Frédéric Bimbot

***Abstract*—We present a general multi-channel source separation framework where additional audio references are available for one (or more) source(s) of a given mixture. Each audio reference is another mixture which is supposed to contain at least one source similar to one of the target sources. Deformations between the sources of interest and their references are modeled in a linear manner using a generic formulation. This is done by adding transformation matrices to an excitation-filter model, hence affecting different axes, namely frequency, dictionary component or time. A nonnegative matrix co-factorization algorithm and a generalized expectation-maximization algorithm are used to estimate the parameters of the model. Different model parameterizations and different combinations of algorithms are tested on music plus voice mixtures guided by music and/or voice references and on professionally-produced music recordings guided by cover references. Our algorithms improve the signal-to-distortion ratio (SDR) of the sources with the lowest intensity by 9 to 15 decibels (dB) with respect to original mixtures.**

***Index Terms*—Source separation, GEM algorithm**

I. Introduction

In audio signal processing, source separation consists in recovering the different audio sources that compose a given observed audio mixture. It has been a hot topic over the past decade and this field of research now offers a wide variety of new possible applications for end-users and professionals. One of those concerns the remastering, restoration and remixing of movie soundtracks or musical recordings. Sound engineers may want to upmix the recordings to a higher number of channels, to remove some sources, to generate a karaoke version, or to substitute some sources by other sources, for instance in order to replace the original soundtrack of a movie with a new one. For these purposes, one needs high source separation quality, which is not yet achievable by blind source separation methods [1]. Taking additional information into account is necessary to improve the separation [2], [3]. In informed source separation methods [4], detailed information about the original sources is transmitted along with the mixture to be separated. Such methods are the ones that provide the best quality but they cannot be applied in the scenario considered hereafter, since the original sources are never observed.

Guided source separation is based on the use of any kind of additional information and has recently been more and more focused on. It is well adapted to scenarios where the original sources are not available but high separation quality is nevertheless required. The additional information can be of different types : spatial and spectral information about the sources [5], [6], language structure [7], visual information [8], information about the recording/mixing conditions [9], musical scores [10]–[13], or user input [14]–[21]. For instance, the user can provide relevant information by drawing the fundamental frequency curve [18], by uttering the same sentence [16], by humming the melody [14], or even by selecting specific areas in the spectrogram of the mixture [17]. On top of this, interactive approaches allow the user to give feedback during the separation [19]–[22].

In this paper, we focus on methods that guide the separation process by a *reference signal* that is similar to one of the target sources [10], [14]–[16], [23]–[26]. Such a framework can be referred to as *reference guided source separation*, and it has recently been used in several scenarios : the restoration of music pieces guided by isolated piano sounds [10], the separation of music and sound effects from speech guided by several versions of the same movie in different languages [23], the separation of musical instruments guided by a multitrack cover version of a song [24], and the denoising of speech guided by the same sentence pronounced by the same speaker [25] or by a different speaker [16]. Symbolic information such as a text [16] or a musical score [12] can also be used to generate reference signals.

Here, we propose a general model for *multi-channel reference guided source separation* that enables the joint use of multiple, multi-channel, deformed reference signals. Our preliminary experiments on music/voice separation [26] showed that in the single-channel case it is relevant to use one reference for both voice and music. Here, we extend this approach to the multi-channel case using a Generalized Expectation-Maximization (GEM) algorithm inspired from [5]. Several initialization procedures are investigated as well as the use of multiple references for each source. Different types of data and references are used to assess the relevance of our approach.

The paper is organized as follows. Section II introduces the general model of reference guided source separation. Section III presents the different algorithms and initialization procedures that are compared in the experiments. Sections IV and V are respectively dedicated to experiments on music/voice separation and on cover-guided music separation.

II. General framework

In this section, we describe the proposed reference guided source separation model. Key concepts are first presented in

Nathan Souviraà-Labastie, Anaik Olivero and Frédéric Bimbot are with PANAMA (Inria/CNRS) project-team at IRISA, Rennes, France. (e-mails: nathan.souviraa-labastie@irisa.fr ; anaik.olivero@inria.fr ; frederic.bimbot@irisa.fr.)

Emmanuel Vincent is with Inria, Villers-lès-Nancy, France. (e-mail: emmanuel.vincent@inria.fr)

The work of Nathan Souviraà-Labastie was supported by MAIA Studio and Bretagne Region scholarship. The work of Anaik Olivero was partly funded by the European Research Council Programme, PLEASE project, under grant ERC-StG-2011-277906.

order to facilitate understanding before we give a detailed description of the model. We also discuss different configurations of our model that can handle different types of available data.

A. Overview of practical scenarios

What we mean by reference signals ranges from different recordings of the true sources to noisy versions of the true sources and also include imitations. The references usually rely on several and very different deformations like time misalignment, time warping, changes of speaker/singer/instrument, additional overlapping sources, equalization, changes of melody/pronunciation, change of recording conditions, and/or pitch shifting. If there is no deformation at all, the reference is then the true source and as already mentioned this is a very restricted scenario. In this sense, we can say that a reference signal is by nature deformed. Hereafter, we consider that there is one target mixture to be separated and several other reference mixtures that contain one (or more) source(s) that are similar enough to the unknown sources of the target mixture.

In this framework, the deformations between the reference signals and the target sources are modeled in a generic linear manner by transformation matrices. Each deformation acts on a specific axis, namely frequency, dictionary component or time. Time-alignment between all these signals is a main concern as it is often needed to provide suitable reference signals. We introduce the notion of multiple multi-channel mixtures to distinguish between phase aligned signals that constitute channels of the same mixture (*e.g.*, different recordings of the true source), and power aligned signals that constitute different mixtures (*e.g.*, approximate imitation). Typically, power alignment is performed by a transformation matrix on the time axis. Examples of initialization and estimation of this time alignment are given in Section IV.

B. Input representation

The observations are M multi-channel audio mixtures $\mathbf{x}^m(t)$ indexed by m and containing I^m channels.

Each mixture $\mathbf{x}^m(t)$ is assumed to be the sum of the spatial images $\mathbf{y}_j(t)$ of one or more sources indexed by $j \in \mathcal{J}_m$:

$$\mathbf{x}^m(t) = \sum_{j \in \mathcal{J}_m} \mathbf{y}_j(t) \text{ with } \mathbf{x}^m(t), \mathbf{y}_j(t) \in \mathbb{R}^{I^m}. \quad (1)$$

In the Short-Time Fourier Transform (STFT) domain, this can be written as

$$\mathbf{x}^m_{fn} = \sum_{j \in \mathcal{J}_m} \mathbf{y}_{j,fn} \text{ with } \mathbf{x}^m_{fn}, \mathbf{y}_{j,fn} \in \mathbb{C}^{I^m}, \quad (2)$$

where $f = 1, ..., F$ and $n = 1, ..., N$ are respectively the frequency and the time indexes of the STFT. We consider that $\mathbf{x}^1(t)$ is the mixture to be separated, and $\mathbf{x}^m(t)$ for $m > 1$ are other mixtures containing the reference signals used to guide the separation process.

We assume that the STFT coefficients of the source spatial images $\mathbf{y}_{j,fn}$ have a zero-mean Gaussian distribution [5] :

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_{\mathbb{C}}(0, v_{j,fn}\mathbf{R}_{j,f}) \quad (3)$$

whose covariance factors into a scalar power spectrum $v_{j,fn} \in \mathbb{R}_+$ and a spatial covariance matrix $\mathbf{R}_{j,f} \in \mathbb{C}^{I^m \times I^m}$.

1) Spatial parameters: The spatial covariance matrices model the spatial characteristics of the sources, such as phase and intensity difference between channels. We only consider time-invariant spatial covariance matrices as the sound sources considered in our experimental scenarios are generally spatially stable over time. $\mathbf{R}_{j,f}$ can be non-uniquely represented as $\mathbf{R}_{j,f} = \mathbf{A}_{j,f}\mathbf{A}^H_{j,f}$ where $\mathbf{A}_{j,f} \in \mathbb{C}^{I^m \times R_j}$ and R_j is the rank of matrices $\mathbf{R}_{j,f}$ and $\mathbf{A}_{j,f}$ [5]. As spectral modeling is the focus of this paper, these aspects are not further detailed.

2) Spectral parameters: The power spectrogram of each source j is denoted as $V_j = [v_{j,fn}]_{fn} \in \mathbb{R}_+^{F \times N}$. Each V_j is split into the product of an excitation spectrogram V_j^e and a filter spectrogram V_j^ϕ. The excitation spectrogram (resp. the filter spectrogram) is decomposed by Nonnegative Matrix Factorization (NMF) into a matrix of spectral patterns $W_j^e \in \mathbb{R}_+^{F \times D^e}$ (resp. $W_j^\phi \in \mathbb{R}_+^{F \times D^\phi}$) and a matrix of temporal activations $H_j^e \in \mathbb{R}_+^{D^e \times N}$ (resp. $H_j^\phi \in \mathbb{R}_+^{D^\phi \times N}$). D^e (resp. D^ϕ) denotes the number of spectral patterns used in the NMF decomposition of the excitation (resp. filter) part. This results in the following decomposition :

$$V_j = V_j^e \odot V_j^\phi = W_j^e H_j^e \odot W_j^\phi H_j^\phi \quad (4)$$

where $\odot$ denotes pointwise multiplication. The four matrices are as follows:

- W_j^e is a spectral dictionary that can be designed as a set of inharmonic, harmonic and/or wideband spectra [5]. Alternatively, such a dictionary can be learned on training data or estimated from the test mixture.
- H_j^e are the corresponding temporal activations which encode, *e.g.*, the musical score in the form of a piano roll [11]–[13], or the f_0 track [18].
- W_j^ϕ is a dictionary of spectral envelopes associated with, *e.g.*, different phonemes in the case of speech [16] or body resonances in the case of a musical instrument [10].
- H_j^ϕ are the corresponding temporal activations which encode, *e.g.*, the phoneme sequence for speech or instrument timbre changes for music such as muted/unmuted trumpet.

C. Proposed model with multiple deformed references

We proposed to consider three different cases for the settings of the matrices W_j^e, H_j^e, W_j^ϕ and H_j^ϕ. As a first case, they can be *fixed* and remain unchanged during the estimation. As a second case, they can be set as *free* parameters, which means that they will be adapted to the corresponding mixture m ($j \in \mathcal{J}^m$) during the estimation process. Finally, as the third case, these matrices can also be *shared* (*i.e.*, jointly estimated) between a given source $j \in \mathcal{J}^m$ and one (or more) reference source(s) $j' \in \mathcal{J}^{m'}$ with $m' \neq m$ (*e.g.*, $m = 1$ as in Sections IV). In this last case, the deformations between sources j and j' are modeled by transformation matrices $T_{jj'}$. We propose to model the sharing of spectral and temporal properties between one source V_j and its references $V_{j'}$ as follows :

1) Transformation matrices for the excitation part: For the excitation part, the transformation matrices are denoted as $T^{fe}_{jj'} \in \mathbb{R}_+^{F' \times F}$, $T^{de}_{jj'} \in \mathbb{R}_+^{D^e \times D^e}$ and $T^{te}_{jj'} \in \mathbb{R}_+^{N' \times N}$.

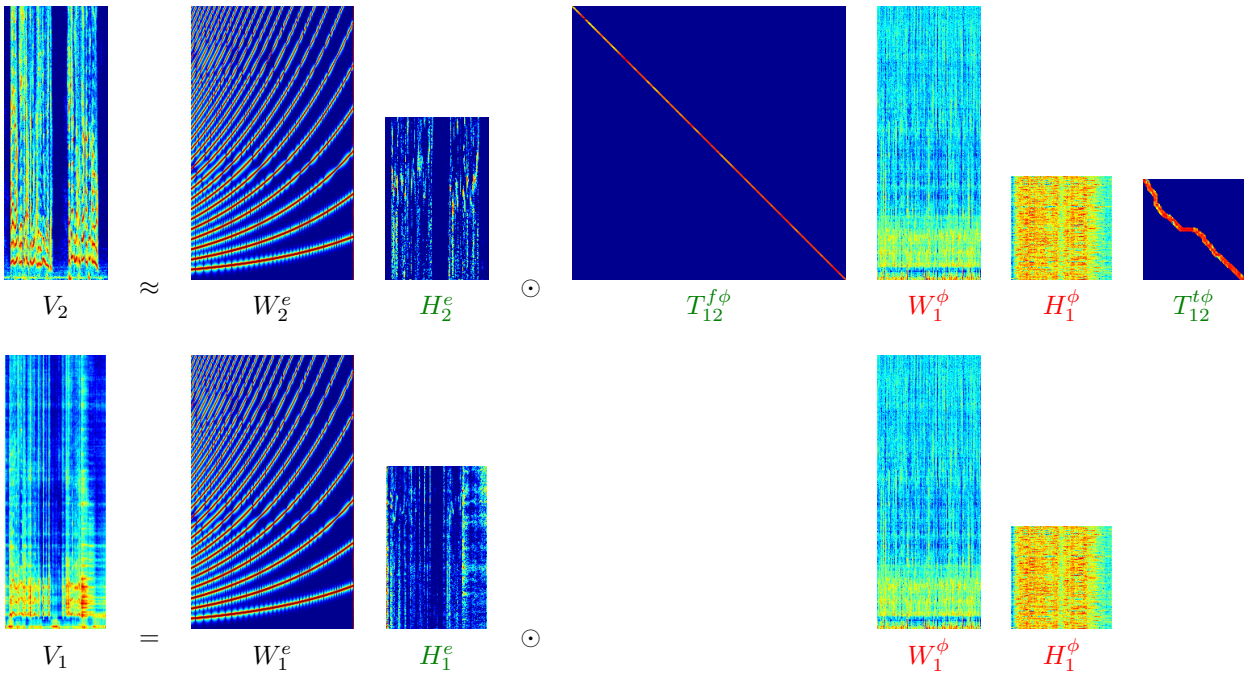

Fig. 1. Example estimated decomposition of the spectral power of a reference mixture ($m' = 2$) containing a single source ($j' = 2$) similar to source $j = 1 \in \mathcal{J}^1$. The excitation parameters are not *shared* whereas the filter follows (10) with $T_{12}^{d\phi}$ set to identity. The parameters of the source of interest ($j = 1$) are also displayed. It can be notice that $X^2 \approx V^2 = V_2$, as there is only one source in the reference mixture.

Depending on the actual deformation between the target source and the reference sources, three different configurations are possible. One may share either the spectral patterns as shown in (5), the temporal activation as shown in (6), or both as shown in (7). This is modeled by one of the three following equations :

$$V_{j'}^e = T_{jj'}^{fe} W_j^e H_{j'}^e \quad (5)$$
$$V_{j'}^e = W_{j'}^e H_j^e T_{jj'}^{te} \quad (6)$$
$$V_{j'}^e = T_{jj'}^{fe} W_j^e T_{jj'}^{de} H_j^e T_{jj'}^{te}. \quad (7)$$

During the estimation process, each transformation matrix can be considered either as a *fixed* or a *free* parameter. In practice, frequency deformations of the excitation $T_{jj'}^{fe}$ can be used to model, *e.g.*, differences of reading speed on analog devices. $T_{jj'}^{te}$ is used to time-align the signal spectra and represents the time warping path between the two signals. $T_{jj'}^{de}$ can be used to model changes in the excitation dictionary, such as pitch shifting[1]. It only appears when the corresponding W_j, H_j are *shared*, otherwise it would be redundant.

2) Transformation matrices for the filter part: For the filter part, the transformation matrices between the target source and the reference sources are denoted as $T_{jj'}^{f\phi} \in \mathbb{R}_+^{F' \times F}$, $T_{jj'}^{d\phi} \in \mathbb{R}_+^{D^\phi \times D^\phi}$ and $T_{jj'}^{t\phi} \in \mathbb{R}_+^{N' \times N}$. In the same way as above, three different configurations are possible, which lead to share either the spectral patterns as shown in (8), the temporal activation as shown in (9), or both as shown in (10) :

$$V_{j'}^\phi = T_{jj'}^{f\phi} W_j^\phi H_{j'}^\phi \quad (8)$$
$$V_{j'}^\phi = W_{j'}^\phi H_j^\phi T_{jj'}^{t\phi} \quad (9)$$
$$V_{j'}^\phi = T_{jj'}^{f\phi} W_j^\phi T_{jj'}^{d\phi} H_j^\phi T_{jj'}^{t\phi}. \quad (10)$$

Similarly to transformation above, the matrices can be either *fixed* or *free*. Frequency deformations of the filter part $T_{jj'}^{f\phi}$ can be used to model, *e.g.*, changes in vocal tract length [16] or a different equalization. $T_{jj'}^{d\phi}$ models changes in the filter dictionary, such as the change of some phonemes in the case of a speaker with a different accent, and it only appears when the corresponding W and H are *shared*. $T_{jj'}^{t\phi}$ models the temporal deformation of the filter, and it is used to time-align the signals.

Figure 1 gives an illustration of a possible use of this model. It corresponds to a speech reference (27) uttered by a different speaker. More details can be found in Section IV-B3a.

D. Comparison with previous approaches

The proposed framework generalizes the state-of-the-art approaches in [10], [16], [26] as they exploit similar models. Our framework can also model the same kind of signals as used in [10], [14], [23]–[25] even if the models can be quite different. Finally, it makes it possible to investigate some new scenarios that have been put forward in [26], like music source separation for a verse guided by another verse.

E. Extensions of our approach

As previously mentioned, we consider here that $j \in \mathcal{J}^1$ and $j' \in \mathcal{J}^{m'}$ with $m' \neq 1$. These notations are supposed

[1]It can be noticed that pitch shifting and reading speed have two different effects, especially for inharmonic sounds.

to represent the classical reference guided source separation scenario. Relaxing this constraint opens the way to more possibilities. Modeling the relationship between sources of the same mixture (*i.e.*, $j, j' \in \mathcal{J}^m$) could be of interest to model delays between sources of the same mixture like a canon in music. Modeling "circular" relationships (*e.g.*, using $T_{jj'}, T_{j'j''}, T_{j''j}$) would allow joint separation of all mixtures. But it requires the use and the estimation of one more matrix. More generally, considering the mixture to be separated as central is a good way to avoid having additional matrices and potential smoothing effects on the sources of interest.

III. Parameter estimation

In this section, we present two methods for parameter estimation in the maximum likelihood (ML) sense. The ML objective can be written as :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda^m \log p(\mathbf{x}^m|\theta) \tag{11}$$

where θ is the set of parameters to be estimated, *i.e.*, the spatial covariance matrices $\mathbf{R}_{j,fn}$, and the matrices W, H and T that are either *free* or *shared*. $\lambda^m \in \mathbb{R}_+$ are weight parameters that can balance potentially different durations or frequency resolutions between mixtures 1 and m, or put more emphasis on the references which are expected to be the most relevant. The reader can refer to [27] for a discussion on their influence on the results.

First, we introduce a multiplicative update (MU) algorithm to deal with single-channel mixtures. Then, a GEM algorithm is used to estimate the parameters in the multi-channel case. Finally, we discuss different initialization procedures.

A. Multiplicative updates for nonnegative matrix partial co-factorization (NMPcF) for the single-channel case

In the single-channel case, maximizing the log-likelihood is equivalent to minimizing the Itakura-Saito divergence [28]:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{m=1}^{M} \lambda^m \sum_{f,n=1}^{F,N} d_{IS}(X^m_{fn}|V^m_{fn}) \tag{12}$$

where $X^m = [|\mathbf{x}^m_{fn}|^2]_{fn}$ and $V^m = \sum_{j\in\mathcal{J}^m} V_j$ are respectively the observed and estimated power spectrograms, and $d_{IS}(a|b) = a/b - \log(a/b) - 1$ is the Itakura-Saito divergence. Other divergences are worth considering but they are not extendable to the multi-channel case whereas Itakura-Saito is. A common way to estimate the parameters is the use of a multiplicative gradient descent approach [28] in which each parameter is updated at each iteration without increasing criterion (12) [29]. The update of one parameter consists in multiplying it by the ratio of the negative and positive parts of the derivative of the criterion with respect to this parameter.

According to their status (*free* or *shared*) different MU can be derived for each parameter. For *free* parameters, (12) leads to the classical MU of NMF. An example of such update is given in (13) for the parameter W^e_j. For *shared* parameters, (12) leads to the MU of NMPcF. An example of such update is given in (14)[2] for the parameter W^e_j.

B. Generalized Expectation-Maximization (GEM) algorithm for the multi-channel case

In the multi-channel case, the spatial information can make the separation clearly more tractable, especially when sources have different directions of arrival. In the case of reference guided source separation, the relevance of multi-channel data remains even though the mixtures have different numbers of channels and even though no assumptions are made on the similarity between directions of arrival of the sources and their references.

Following the general framework in [5], we introduce R_j independent Gaussian random variables $s_{jr,fn}$ $(r = 1, ..., R_j)$ distributed as $s_{jr,fn} \sim \mathcal{N}_{\mathbb{C}}(0, v_{j,fn})$ for every source j and time-frequency bin (f, n). An additive isotropic noise source $\mathbf{b}^m_{fn}$ of diagonal covariance $\Sigma_{\mathbf{b}^m_{fn}} = \sigma^2_f \mathbf{I}_{I^m} \in \mathbb{C}^{I^m \times I^m}$ is also added for each mixture m. With these changes, (2) becomes :

$$\mathbf{x}^m_{fn} = \mathbf{A}^m_{fn} \mathbf{s}^m_{fn} + \mathbf{b}^m_{fn} \tag{17}$$

where $\mathbf{A}^m_{fn} \in \mathbb{C}^{I^m \times R^m}$ (resp. $\mathbf{s}^m_{fn} \in \mathbb{C}^{R^m \times I^m}$) results from the concatenation ($R^m = \sum_{j\in\mathcal{J}^m} R_j$) of the mixing matrices $\mathbf{A}_{j,fn}$ (resp. of all the sub-sources $s_{jr,fn}$) of all the sources $j \in \mathcal{J}^m$.

EM is a natural algorithm to handle such a parameter estimation in the ML sense in the presence of *observed* data $\mathbf{X} = \{\mathbf{X}^m\}_m = \left\{x^m_{fn}\right\}_{m,fn}$ and *unobserved* data $\mathbf{S} = \{\mathbf{S}^m\}_m = \left\{s^m_{fn}\right\}_{m,fn}$, that form a complete set $\mathbf{Z} = \{\mathbf{X}, \mathbf{S}\}$. The algorithm proceeds by alternating an E-step that computes the expected value of the complete-data log-likelihood $\mathbb{E}_{\mathbf{Z}|\theta^c}\left[\log p(\mathbf{Z}|\theta)\right] \triangleq Q(\theta, \theta^c)$ given the observation and the current set of parameters θ^c, and an M-step that chooses a θ that maximizes the quantity $Q(\theta, \theta^c)$. In the case of GEM, the M-step only seeks to find a θ that increases Q. A detailed derivation is given in Appendix A. The quantity Q can be written up to a constant as :

$$\begin{aligned} Q(\theta, \theta^c) \overset{c}{=} - \sum_{mfn} \frac{\lambda^m}{\sigma^2_f} \operatorname{tr} \Big[& \mathbf{R}_{\mathbf{x}^m_{fn}} - \mathbf{A}^m_f \mathbf{R}_{\mathbf{xs}^m_{fn}}{}^H \\ & - \mathbf{R}_{\mathbf{xs}^m_{fn}} \mathbf{A}^{mH}_f + \mathbf{A}^m_f \mathbf{R}_{\mathbf{s}^m_{fn}} \mathbf{A}^{mH}_f \Big] \\ & - \sum_{m,j\in\mathcal{J}^m,fn} \lambda^m R_j d_{IS}(\xi_{j,fn}|v_{j,fn}), \end{aligned} \tag{18}$$

with: $\mathbf{R}_{\mathbf{x}^m_{fn}} \triangleq \hat{\mathbf{R}}_{\mathbf{x}^m_{fn}} = \hat{\mathbb{E}}[\mathbf{x}^m_{fn} \mathbf{x}^m_{fn}{}^H]$, $\mathbf{R}_{\mathbf{xs}^m_{fn}} \triangleq \hat{\mathbb{E}}[\mathbf{x}^m_{fn} \mathbf{s}^m_{fn}{}^H]$, $\mathbf{R}_{\mathbf{s}^m_{fn}} \triangleq \hat{\mathbb{E}}[\mathbf{s}^m_{fn} \mathbf{s}^m_{fn}{}^H]$ and $\xi_{j,fn} \triangleq \frac{1}{R_j} \sum_{r=1}^{R_j} \hat{\mathbb{E}}\left[|s_{jr,fn}|^2\right]$.

Starting from (18), one can demonstrate that $\mathbb{T}(\mathbf{X}, \mathbf{S}) = \left\{\mathbf{R}_{\mathbf{x}^m_{fn}}, \mathbf{R}_{\mathbf{xs}^m_{fn}}, \mathbf{R}_{\mathbf{s}^m_{fn}}\right\}_{m,f,n}$ is the set of *natural (sufficient) statistics* [30] for $\mathbf{Z}$. This leads to the following two steps of our GEM algorithm.

[2] In (14) and (16), we assume that $V^e_{j'}$ and $V^\phi_{j'}$ follow models (7) and (10). In practical scenarios, the number of *shared* parameters and transformation matrices will generally be smaller, as exemplified in Sections IV and V.

MU-free :

$$W_j^e \leftarrow W_j^e \odot \frac{[V_j^\phi \odot V^{m\cdot[-2]} \odot X^m][H_j^e]^T}{[V_j^\phi \odot V^{m\cdot[-1]}][H_j^e]^T} \tag{13}$$

MU-shared :

$$W_j^e \leftarrow W_j^e \odot \frac{\lambda^m[V_j^\phi \odot V^{m\cdot[-2]} \odot X^m][H_j^e]^T + \sum_{j'} \lambda^{m'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V^{m'\cdot[-2]} \odot X^{m'}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T}{\lambda^m[V_j^\phi \odot V^{m\cdot[-1]}][H_j^e]^T + \sum_{j'} \lambda^{m'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V^{m'\cdot[-1]}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T} \tag{14}$$

EM-free :

$$W_j^e \leftarrow W_j^e \odot \frac{[V_j^\phi \odot V_j^{\cdot[-2]} \odot \hat{\Xi}_j][H_j^e]^T}{[V_j^\phi \odot V_j^{\cdot[-1]}][H_j^e]^T} \tag{15}$$

EM-shared :

$$W_j^e \leftarrow W_j^e \odot \frac{\lambda^m R_j[V_j^\phi \odot V_j^{\cdot[-2]} \odot \hat{\Xi}_j][H_j^e]^T + \sum_{j'} \lambda^{m'} R_{j'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V_{j'}^{\cdot[-2]} \odot \hat{\Xi}_{j'}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T}{\lambda^m R_j[V_j^\phi \odot V_j^{\cdot[-1]}][H_j^e]^T + \sum_{j'} \lambda^{m'} R_{j'}[T_{jj'}^{fe}]^T[V_{j'}^\phi \odot V_{j'}^{\cdot[-1]}][T_{jj'}^{de} H_j^e T_{jj'}^{te}]^T} \tag{16}$$

1) E-step: This step consists in computing the conditional expectations of the natural statistics given θ^c:

$$\hat{\mathbf{R}}_{\mathbf{s}_{fn}^m} = \mathbf{\Omega}_{\mathbf{s}_{fn}^m} \hat{\mathbf{R}}_{\mathbf{x}_{fn}^m} \mathbf{\Omega}_{\mathbf{s}_{fn}^m}{}^H + (\mathbf{I}_R - \mathbf{\Omega}_{\mathbf{s}_{fn}^m} \mathbf{A}_f^m)\mathbf{\Sigma}_{\mathbf{s}_{fn}^m} \tag{19}$$

$$\hat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{fn}^m} = \hat{\mathbf{R}}_{\mathbf{x}_{fn}^m} \mathbf{\Omega}_{\mathbf{s}_{fn}^m}{}^H \in \mathbb{C}^{I^m \times R^m} \tag{20}$$

with :

$$\mathbf{\Sigma}_{\mathbf{s}_{fn}^m} = diag([\phi_{r,fn}]_{r=1}^{R^m}) \in \mathbb{R}_+^{R^m \times R^m} \tag{21}$$

$$\mathbf{\Sigma}_{\mathbf{x}_{fn}^m} = \mathbf{A}_f^m \mathbf{\Sigma}_{\mathbf{s}_{fn}^m} \mathbf{A}_f^{mH} + \mathbf{\Sigma}_{\mathbf{b}_{fn}^m} \in \mathbb{C}^{I^m \times I^m} \tag{22}$$

$$\mathbf{\Omega}_{\mathbf{s}_{fn}^m} = \mathbf{\Sigma}_{\mathbf{s}_{fn}^m} \mathbf{A}_f^{mH} \mathbf{\Sigma}_{\mathbf{x}_{fn}^m}{}^{-1} \in \mathbb{C}^{R^m \times I^m} \tag{23}$$

where $\phi_{r,fn} = v_{j,fn}$ if $r \in \mathcal{R}_j$ (*i.e.*, r is a sub-source of j).

2) M-step: The *free* parameters that compose the set θ are here updated in order to increase the quantity Q. The update for the spatial parameters is [5] :

$$\mathbf{A}_{j,f} = \left[\sum_n \hat{\mathbf{R}}_{\mathbf{x}\mathbf{s}_{fn}^m}\right] \left[\sum_n \hat{\mathbf{R}}_{\mathbf{s}_{fn}^m}\right]^{-1}. \tag{24}$$

If none of the parameters are *shared*, the resulting GEM algorithm processes the different mixtures separately, and behaves as in [5]. The sharing of spectral parameters induces a single change in the algorithm routines that occurs during the M-step updates of these *shared* spectral parameters. Examples of MU are given for *free* parameters in (15) and for *shared* parameters in (16)[2]. They generalize the update (30) in [5] with $\hat{\Xi}_j = [\hat{\xi}_{j,fn}]_{fn} \in \mathbb{R}_+^{F \times N}$ where $\hat{\xi}_{j,fn} = \frac{1}{R_j} \sum_{r=1}^{R_j} \hat{\mathbf{R}}_{\mathbf{s}_{fn}^m}(r,r)$.

C. Parameter initialization

The results of both MU and EM depend on initialization. With respect to blind source separation, reference guided separation provides better initial values for the parameters W and H taking advantage of the provided references. The other parameters, *i.e.*, the transformation matrices T, are beforehand roughly estimated (see Section IV). For instance, we can use MU to minimize the following criterion :

$$\hat{\theta}_{\text{ref}} = \underset{\theta_{\text{ref}}}{\text{argmin}} \sum_{m=2}^{M} \lambda^m \sum_{f,n=1}^{F,N} d_{IS}(X_{fn}^m | V_{fn}^m) \tag{25}$$

where θ_{ref} is the set of W and H parameters that occur in the reference signals. This is especially efficient when there is a single dominant source in each reference signal. At the end of this stage, only the parameters of the mixture to be separated that are not shared with any reference signals remain weakly initialized if no prior information about them is available.

In the experiments, we will distinguish the following successive initialization and algorithmic stages :

- *Init* : for all sources, we define the status (*i.e.*, *fixed*, *free* or *shared*) of their spectral parameters (*i.e.*, W, H and T) and we initialize them in the best way according to prior information. The exact initialization is specified below in Section IV and V depending on the considered scenario.
- *NMF* : the *shared* and *free* W, H of the reference mixtures are updated using MU, *i.e.*, the method described in this Section III-C,
- *NMPcF* : we apply on all mixtures the algorithm described in Section III-A,
- *GEM* : we apply on all mixtures the algorithm described in Section III-B.

During the experiments, different combinations of these four stages are tried in the above order. In all cases, the final source estimates are obtained using an adaptive Wiener filter $\hat{\mathbf{s}}_{fn}^m = \mathbf{\Omega}_{\mathbf{s}_{fn}^m} \mathbf{x}_{fn}^m$ and multiplied by the structured A_f^m to obtain the corresponding spatial images $\mathbf{y}_{j,fn}^m$.

IV. Voice/music separation

In this section, we describe in details one use case of the source separation framework with deformed reference described above. We target the separation of speech and music from old recorded movies and TV series. Speech and/or music references are used to guide the separation.

After briefly describing the data, we recall how speech and music references are modeled in our proposed framework. We consider two different models for the music references depending on whether they are aligned a priori or not. We conduct experiments that compare these two models as well as different initialization procedures. Finally, we investigate the

use of several references for a single source with the objective of making the separation more robust.

A. Data

The musical samples and the corresponding references are obtained using the MODIS audio motif discovery software in [31]. This software aims at clustering the segments of a long audio stream (here movies or TV series) that are similar enough according to a threshold. It is based on seeded discovery and template matching [32], but on a more fundamental level the audio segments are compared using a segmental variant of dynamic time warping (DTW) and common features. As long as it allows the discovery of non-exact repetitions, the discovered references are distorted compared to the source of interest (rhythm changes, fade in) and also contain additional sources (mainly sound effects).

Speech examples are taken from the database in [33] in which 16 different speakers uttered the same 238 sentences. We kept 4 musical examples and 4 sentences (two female and two male speakers) and mixed them at two different voice-to-music ratios : -6 dB (music as foreground and voice as background), and 12 dB (voice as foreground and music as background). These levels are close to those effectively observed in movies and TV series. We mixed such examples ourselves in order to obtain objective measures for the evaluation and to compare our estimated sources with the original ones using [34]. Combining those parameters leads to 32 original mixtures X^1. For each mixture to be separated, we have one or more deformed music reference(s) (other discovered versions of the same music excerpt), and one or more deformed speech reference(s) (same sentence uttered by different speakers). The original mixtures and the references are about eight seconds long and they are sampled at 16 kHz. Some examples are available online[3].

B. Tested models

In the different setups reported here, the speech sources are numbered as $j = 1$ or 2, the music sources as $j = 3$ or 4, and the other sources and background noise as $j = 5$ or 6. *Fixed* variables are in black ($W_1^e, W_2^e, W_3^e, W_4^e, T_{34}^{t\phi}$). *Free* variables are in green (H_1^e, H_2^e, $T_{12}^{f\phi}$, $T_{12}^{d\phi}$, $T_{12}^{t\phi}$, T_{34}^{te}, W_5, H_5, W_6, H_6). *Shared* variables are in red or magenta (W_1^ϕ, H_1^ϕ, H_3^e, W_3^ϕ, H_3^ϕ). Note that, in one particular setup, H_3^e, W_3^ϕ and H_3^ϕ are *free*. The *fixed* matrices T set to identity are removed from the notations.

1) Signal to be separated: The first signal is the mixture to be separated. It is composed of speech V_1, music V_3 and noise V_5 :

$$\begin{aligned} V^1 &= V_1 + V_3 + V_5 \\ &= W_1^e H_1^e \odot W_1^\phi H_1^\phi + W_3^e H_3^e \odot W_3^\phi H_3^\phi + W_5 H_5 \end{aligned} \quad (26)$$

2) Speech reference: The second mixture is composed of the speech reference V_2 alone :

$$V^2 = V_2 = W_2^e H_2^e \odot T_{12}^{f\phi} W_1^\phi T_{12}^{d\phi} H_1^\phi T_{12}^{t\phi} \quad (27)$$

[3]http://speech-demos.gforge.inria.fr/source_separation/taslp2015/index.html

During the *NMPcF* and/or *GEM* stages, H_1^e and H_2^e are estimated separately to model the different intonations and pitches between the speakers. Conversely, the filter matrices W_1^ϕ and H_1^ϕ are jointly estimated to model similar phonetic content, as the two speech signals are composed of the same phonemes. $T_{12}^{t\phi}$ models the time alignment between the two utterances. $T_{12}^{f\phi}$ is constrained to be diagonal and it models both the equalization and the speaker's difference. $T_{12}^{d\phi}$ is also used to model the speaker difference and its initialization is discussed in Section IV-F. This model is similar to the one used in [16], where a time-invariant filter is used to model the frequency deformations of V_2, but here we consider in addition a transformation matrix $T_{12}^{d\phi}$ for the filter dictionary. Additional speech references are treated in the same way.

3) Music reference: The third signal is composed of the music reference V_4, that is similar to V_3, and of some noise V_6. We consider two different models for this signal.

a) non phase aligned reference: This first model represents the reference signal as a third observed mixture V^3 :

$$V^3 = V_4 + V_6 = W_4^e H_3^e T_{34}^{te} \odot W_3^\phi H_3^\phi T_{34}^{t\phi} + W_6 H_6 \quad (28)$$

T_{34}^{te} and $T_{34}^{t\phi}$ model the alignment of short-term spectra between the two music examples.

b) phase aligned reference: In this second model, the music reference is aligned at the signal level with the mixture to be separated. This is achieved by windowing X^1 into time frames, aligning each time frame with X^2 by means of time-delay computation using the Generalized Cross Correlation with PHase Amplitude Transform (GCC-PHAT) algorithm [35], and reconstructing a time-aligned reference signal $\tilde{X}^2$ by overlap-and-add. Once aligned, it is considered within our general framework as a second channel of the first mixture ($I^1 = 2$). As a consequence, V^3 is not used in this model, and V_3 becomes a single music source with a *free* spatial parameter $A_3 \in \mathbb{C}^{2\times 2}$ encoding the amplitude and phase differences between the target and the reference channel of the first mixture. H_3^e, W_3^ϕ, and H_3^ϕ are *free* parameters. As the other sources belong to one channel only, their spatial parameters are *fixed* to $A_1 = A_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and $A_6 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

C. Parameter initialization

We here give details about the *Init* and *NMF* stages previously mentioned in Section III-C. The weight parameter $\lambda^{m'}$ is set to $\frac{NF}{N'F'}$. The *fixed* excitation spectral patterns W_j^e for $j = 1, 2, 3, 4$ are a set of harmonic components computed as in [5]. We initialize the synchronization matrices $T_{12}^{t\phi}$, T_{34}^{te}, and $T_{34}^{t\phi}$ with Dynamic Time Warping (DTW) [36] matrices computed on MFCC vectors [37] for speech sources and on chroma vectors for music sources. Following [16], we allow the temporal path to vary within an enlarged region around the estimated DTW path. Since the data are deformed and noisy (especially for music), we weight this enlarged path by coefficients of the similarity matrix (from which the DTW is obtained), in order to avoid obvious initialization errors. As we work with MU, let us remind that zeros in the parameters remain unchanged over the iterations. We invite the reader

to refer to [16] for details on this strategy. The spectral transformation matrix $T_{12}^{f\phi}$ is initialized as the identity matrix. The others matrices (H_1^e, H_2^e, W_5, H_5, W_6, H_6, W_1^ϕ, H_1^ϕ, H_3^e, W_3^ϕ, H_3^ϕ) are initialized with random values. In the phase aligned music reference case, we proceed with the same initialization as for the previous setup, but without any T^{te} and $T^{t\phi}$ matrices for music. The music spatial parameter is initialized to $A_3 \approx \begin{pmatrix} 1 & -0.25 \\ 1 & 0.25 \end{pmatrix}$, where the first column accounts for the fact that the two channels are expected to be time and amplitude aligned and the second column accounts for residual differences.

The *NMF* stage can then be applied separately on the reference mixtures (27) and (28) (unaligned references), where the *shared* matrices (W_1^ϕ, H_1^ϕ, H_3^e, W_3^ϕ, H_3^ϕ) and the free parameters (H_2^e, W_6, H_6) are updated whereas matrices $T_{12}^{f\phi}$, $T_{12}^{d\phi}$, $T_{12}^{t\phi}$, T_{34}^{te}, and $T_{34}^{t\phi}$ are not. In the phase aligned music reference case, the parameters H_3^e, W_3^ϕ, H_3^ϕ, W_6, and H_6 are updated to fit the reference signal that is already phase aligned with the signal to be separated. In both cases, W_6 and H_6 are set once again to random values before applying *NMPcF* and/or *GEM*.

D. Algorithm combination

As a first experiment, we evaluate the effect of NMF initialization in the case of a single non phase aligned music reference and no speech reference. The number of iterations is set to 10 for the *NMF* and *NMPcF* stages, and to 100 for the *GEM* stage that is known to require more iterations. The separation performanceresults are evaluated in terms of signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifacts ratio (SAR) [34].

The results are summarized in the top part of Table I. The best SDRs are indicated in bold for each column's part (delimited by double lines). A notable improvement (at least 2.5 dB) is observed when *NMF* and *NMPcF* are combined, as compared to using *NMPcF* alone.

E. Model comparison for music reference

As a second experiment, we evaluate the effect of phase alignment for the music reference. The results are shown in the bottom part of the Table I. Comparing the configurations chosen for the third and fourth lines with the first line shows that *GEM* decreases the separation performance if it is used directly after *NMF* or *NMPcF* alone. This highlights that the proper functioning of the *GEM* depends on the two previous steps. The best results are obtained when the signals are phase aligned and both *NMF* and *NMPcF* are used before *GEM*. In that case, the improvements compared to the non phase aligned case are marginal when the music is as foreground, but significant when the music is as background.

As a *GEM* iteration is in the order of ten times longer than a *NMPcF* iteration, the relevance of this costly last step can be discussed. Given the marginal improvements when music is in the foreground, additional *GEM* iterations are not necessary for this voice-to-music ratio. Conversely, when music is in the background, *GEM* increases the music SDR by 2.4 dB. This result is also greater by 1.3 dB compared to approaches that use multiple references (see Table II).

F. Multiple references for a single source

Experiments on the effect of using reference signals for different sources have been conducted in [26]. We here conduct complementary experiments on the influence of the number of reference signals per source, *i.e.*, several j' for a single j. The number of speech (resp. music) references grows from 0 to 3 (resp. 2). Table II gathers all the results. The best SDRs are indicated in bold.

It can be emphasized that the use of multiple speech references leads to better result, especially when speech is in the background (in the order of 0.5 dB). Conversely, the use of two music references tends to smaller or equal results. This can be explained by the fact that the considered music references contain additional sources supposed to be taken into account by the matrices W_6 and H_6 in (28), and that leads to a more complicated situation for the algorithm. But for some particular examples, the second music reference leads to better results. Overall, the addition of more than one reference seems to improve separation when the new references carry complementary information.

V. Cover guided music separation

This second experimental part focuses on the task of professionally-produced music separation guided by covers [24]. A cover song is a replica of an original song with some differences due for instance to artist interpretation, singer/instrument changes, or new song structure. Such covers can be easily found, and they are usually close to the original song making them interesting for separation. As it provides high quality separation, such demixing enables the edition of the song by end-users (*e.g.*, for active listening) or professional users (*e.g.*, for upmixing).

Here we use multitrack recordings of cover songs to guide the separation. Each track is used as a reference for one corresponding source, so the number of tracks is the same as the number of sources to be separated. In [24], the multitrack cover signals are only used to initialize the source parameters W and H. Here, these parameters are shared between the source and the reference hence the reference signals are used during the estimation stage too. Deformations are modeled in various ways using the general framework introduced in Section II.

A. Data and settings

In order to compare our results, we used the same data set and settings as in [24]. Both original and cover multitracks are available in order to evaluate the separation. They are also used in the mirror configuration, *i.e.*, considering the original as the reference and vice versa. Here, we make an exhaustive list of settings that differ from [24] and refer the reader to [24] for other common details.

The 30 second examples are chosen in a different way as in [24] and are typically composed of half of a verse and half

		-6 dB voice-to-music ratio						12 dB voice-to-music ratio					
		voice			music			voice			music		
		SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
No phase alignment	*Init + NMPcF*	-0.6	2.3	-2.1	5.9	11.7	8.3	3.6	4.9	29.2	-6.8	6.8	-5.5
	Init + NMF + NMPcF	**2.1**	2.9	8.1	**9.2**	11.6	17.7	6.0	8.7	24.6	**0.5**	2.7	3.9
Phase alignment	*Init + NMF + GEM*	0.0	5.7	-3.3	1.8	-1.3	16.3	4.1	4.3	10.7	-8.7	-9.3	4.6
	Init + NMPcF + GEM	-1.1	3.9	-1.0	5.3	11.6	8.4	3.2	3.8	27.9	-7.3	6.7	-6.0
	Init + NMF + NMPcF + GEM	**2.2**	3.7	7.5	**9.8**	11.4	17.7	**7.6**	13.0	21.6	**2.9**	4.0	10.0

TABLE I
AVERAGE VOICE/MUSIC SEPARATION PERFORMANCE (DB) FOR DIFFERENT COMBINATIONS OF ALGORITHMIC STAGES IN THE CASE OF ONE MUSIC REFERENCE AND NO VOICE REFERENCE.

Number of speech references	Number of music references	-6 dB voice-to-music ratio						12 dB voice-to-music ratio					
		voice			music			voice			music		
		SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
1	0	2.1	5.9	3.7	7.7	11.9	13.8	8.7	11.3	19.7	**-2.2**	3.5	-1.5
2	0	2.3	6.1	3.9	7.9	12.4	13.3	8.6	11.1	19.9	-2.5	3.8	-1.9
3	0	**2.8**	5.7	4.7	**8.3**	12.6	13.6	**9.2**	11.7	20.6	**-2.2**	4.1	-1.6
0	1	2.1	2.9	8.1	**9.2**	11.6	17.7	6.0	8.7	24.6	0.5	2.7	3.9
1	1	4.6	6.0	9.9	8.0	9.6	18.9	**13.3**	14.5	26.2	**1.6**	3.4	6.7
2	1	4.9	6.2	10.2	8.5	10.4	19.4	12.2	13.5	25.5	0.9	3.4	6.5
3	1	**5.0**	6.3	10.5	8.6	10.5	19.4	12.1	13.4	25.4	**1.6**	3.4	6.3
0	2	1.5	3.0	4.9	8.4	11.8	14.4	4.8	7.5	25.3	-2.2	3.9	-1.0
1	2	4.1	6.1	8.5	8.1	11.0	16.2	**10.4**	12.2	26.2	-0.6	3.4	1.2
2	2	**4.6**	6.3	9.4	8.3	11.1	16.5	10.3	12.2	26.2	-0.9	3.6	1.3
3	2	**4.6**	6.2	9.5	**8.5**	11.3	16.8	10.0	11.8	25.7	**-0.5**	3.7	1.2

TABLE II
AVERAGE VOICE/MUSIC SEPARATION PERFORMANCE (DB) FOR DIFFERENT NUMBERS OF SPEECH AND MUSIC REFERENCES. SPEECH REFERENCES ARE UTTERED BY DIFFERENT SPEAKERS THAT HAVE THE SAME GENDER AS IN THE MIXTURE TO BE SEPARATED. THE MUSIC REFERENCE IS NOT PHASE ALIGNED, AND ONLY 10 ITERATIONS OF *NMF* AND *NMPcF* ARE USED.

Title	**Track names**
I Will Survive	Bass, Brass, Drums, Electric Guitar, Strings, Vocal.
Pride and Joy	Bass, Drums, Electric Guitar, Vocal.
Rocket Man	Bass, Choirs, Drums, Others, Piano, Vocal.
Walk this Way	Bass, Drums, Electric Guitar, Vocal.

TABLE III
COVER MULTITRACK DATASET

of a chorus. The considered tracks of four songs are listed in Table III. There is no second electric guitar for the song "Walk this Way" as it does not appear in the example that we selected. We use 50 iterations for *NMF* and *NMPcF* steps instead of 500 [24]. The number of components D is kept to 50. For the sake of clarity, only the single-channel case is investigated.

B. Tested models

In this scenario, the reference signals are the different tracks of the cover and each reference signal is related to a single source in the mixture to be separated (the original song). The power spectrum of each reference signal is modeled as $V_j = W_j H_j$ and its parameters are initialized using the *Init* and *NMF* stages in all the different settings reported in Table IV.

Conversely, different settings are considered for the power spectrum model of the sources to be separated: WH, $T^f WH$, and $WT^d H$ (see Table IV). The general framework introduced in Section II is then used by inverting j and j' in (7) for instance. This is because the initialization of the T^f and T^d matrices is weak and would have disturbed the *NMF* stage.

C. Initialization

Parameters W and H are randomly initialized before being updated to fit the reference signal. Here, this is done by the so-called *Init* and *NMF* steps similarly to what was done in [24]. When used, T^f is initialized as an identity matrix. Along the same lines, we tested several initializations of T^d starting from the identity matrix and changing the weight of the off-diagonal coefficients. The best initial value was the sum of an identity matrix and a random matrix drawn from a rectified Gaussian distribution.

D. Results

Separation performance is evaluated in terms of signal-to-distortion ratio improvement (SDRI) that is the difference between the output SDR [34] and the input SDR. The input SDR is defined as the power ratio between a source to be estimated and the mixture to be separated. The samples that we selected lead to an input SDR of the same order (-8.44 dB instead of -7.60 dB in [24]). The results are summarized in Table IV.

First, we reproduced the experiments in [24] with the differences previously presented. A similar SDRI mean is obtained (8.74 dB instead of 8.98 dB) in the case when the parameters are not shared. Compared to the initialization (10.06 dB), this configuration leads in fact to a decrease of the average SDRI. This can be explained by the high similarity between the covers and the original tracks. Conversely, sharing the parameters during the final estimation guarantees not to

Source model	**Instruments** Number of tracks Input SDR	**Average** 20 -8.44	**Bass** 4 -7.42	**Drums** 4 -7.17	**Guitar** 3 -9.98	**Vocal** 4 -4.18	**Choirs** 1 -12.34	**Others** 1 -9.75	**Piano** 1 -12.48	**Brass** 1 -18.64	**Strings** 1 -10.55
WH	*Init + NMF*	10.06	9.33	9.02	9.71	9.60	13.70	9.79	10.80	16.25	9.80
WH	*Init + NMF* + NMF of [24]	8.74	6.94	8.95	8.53	8.14	11.36	9.66	10.51	11.54	10.11
WH	*Init + NMF + NMPcF*	10.27	**9.26**	9.28	9.82	10.24	**13.23**	10.27	11.11	15.67	10.62
T^fWH	*Init + NMF + NMPcF*	10.09	8.79	9.22	9.88	9.78	12.29	**10.47**	**12.94**	14.66	10.65
WT^dH	*Init + NMF + NMPcF*	10.64	9.23	**9.94**	**10.80**	**10.73**	13.01	10.07	11.36	**15.80**	**10.61**
Best	*Init + NMF + NMPcF*	10.85	9.42	10.25	10.11	10.79	14.42	10.05	11.47	18.09	10.83
WH	*Init* + Oracle	15.48	14.48	15.35	15.36	15.29	16.21	13.27	16.22	21.58	15.74

TABLE IV
AVERAGE SDRI (DB) FOR THE SEPARATION OF MUSIC RECORDINGS USING MULTITRACK COVERS AS REFERENCES.

draw away too much from the starting point while getting closer to a solution that fits better to the original tracks. In our case, a marginal improvement is observed (10.27 dB average SDRI). This small improvement is not surprising as we considered no deformation between the references and the sources that we estimate.

More appropriate models that use deformations (T^f or T^d) have been tried with promising results. Better overall results (10.64 dB) are obtained when using a T^d in the source model, whereas T^f slightly decreases the results. Moreover, the improvement is not uniformly observed, and some sources are more enhanced using T^f instead of T^d.

We conduct a final experiment (**Best**) where, for each source, the best source model (values in bold font) is chosen. In that case, we observe an overall increase of the performance (10.85 dB), hence showing the potential of using suitable deformation models for specific sources. However, as we selected the different models based on their optimal results, this experiment is of course not representative of a realistic scenario.

VI. CONCLUSION

In this paper, we have presented a general framework for using audio information in order to separate a given mixture. This model is general enough to take different types of audio references into account and to accomodate for their possible deformations in the frequency domain and/or in the temporal domain. We have provided extensive experiments on two realistic scenarios : voice and music separation in the context of movie soundtracks, and cover-guided music separation.

Our experiments show that the use of reference for a given source improves the general sound quality of the estimated source (from 9 to 15 dB). Different temporal alignment methods appear to be adapted to different situations and signals. Moreover, our experiments show that having at least one reference per source is of prime importance.

A more general perspective of this study will be the design of some automatic processes that choose the best configuration of our general model for a given mixture. Our model can also be improved by adding well-chosen constraints on the parameters. For instance, smoothness constraints on the spectral transformation matrices $T^{f\phi}_{j'j}$ should help the model to derive a more relevant spectral deformation between the target sources and the references.

ACKNOWLEDGMENT

The authors would like to thank MAIA Studio for their sound engineering expertise and for partly funding this work.

REFERENCES

[1] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

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

[3] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, May 2014.

[4] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1699–1712, 2013.

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

[6] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *Show & Tell IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italie, May 2014.

[7] G. J. Mysore and P. Smaragdis, "A non-negative approach to language informed speech separation," in *Proc. 10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel-Aviv, Israel, 2012, pp. 356–363.

[8] A. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 358–371, Aug. 2010.

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

[10] U. Simsekli, Y. K. Yilmaz, and A. T. Cemgil, "Score guided audio restoration via generalised coupled tensor factorisation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 5369–5372.

[11] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, Oct 2012, pp. 277–282.

[12] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 888–891.

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

APPENDIX A

Here we give a detailed derivation of the expected value of the complete-data log-likelihood (18).

$$
\begin{aligned}
Q(\theta,\theta^c) &\triangleq \mathbb{E}_{\mathbf{Z}|\theta^c}\left[\log p(\mathbf{Z}|\theta)\right] = \sum_m \lambda^m \mathbb{E}_{\mathbf{Z}|\theta^c}[\log p(\mathbf{X}^m,\mathbf{S}^m|\theta)] \\
&= \sum_m \lambda^m \mathbb{E}_{\mathbf{X}^m,\mathbf{S}^m|\theta^c}[\log p(\mathbf{X}^m|\mathbf{S}^m)] + \sum_m \lambda^m \mathbb{E}_{\mathbf{S}^m|\theta^c}[\log p(\mathbf{S}^m|\theta)] \\
&= \sum_{m,f,n} \lambda^m \mathbb{E}_{\mathbf{x}_{fn}^m,\mathbf{s}_{fn}^m|\theta^c}\left[\log \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{fn}^m|\mathbf{A}_{fn}^m\mathbf{s}_{fn}^m,\mathbf{\Sigma}_{\mathbf{b}_{fn}^m}\right)\right] + \sum_{m,j\in\mathcal{J}^m,f,n} \lambda^m \sum_{r=1}^{R_j} \mathbb{E}_{s_{jr,fn}|\theta^c}\left[\log \mathcal{N}_{\mathbb{C}}\left(s_{jr,fn}|0,v_{j,fn}\right)\right] \\
&\overset{c}{=} -\sum_{mfn} \frac{\lambda^m}{\sigma_f^2} \mathrm{tr}\left[\mathbf{R}_{\mathbf{x}_{fn}^m} - \mathbf{A}_{fn}^m \mathbf{R}_{\mathbf{x}\mathbf{s}_{fn}^m}{}^H - \mathbf{R}_{\mathbf{x}\mathbf{s}_{fn}^m}\mathbf{A}_{fn}^m{}^H + \mathbf{A}_{fn}^m \mathbf{R}_{\mathbf{s}_{fn}^m}\mathbf{A}_{fn}^m{}^H\right] - \sum_{m,j\in\mathcal{J}^m,fn} \lambda^m R_j d_{IS}(\xi_{j,fn}|v_{j,fn})
\end{aligned}
$$

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

[15] D. FitzGerald, "User assisted separation using tensor factorisations," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 2412–2416.

[16] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2013)*, Southampton, United Kingdom, Sept. 2013, pp. 1–6.

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

[18] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. 10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel-Aviv, Israel, 2012, pp. 438–445.

[19] B. Fuentes, R. Badeau, and G. Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 2654–2658.

[20] N. J. Bryan and G. J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates." in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 883–887.

[21] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italie, May 2014, pp. 1567–1571.

[22] N. J. Bryan, G. J. Mysore, and G. Wang, "ISSE: an interactive source separation editor," in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 257–266.

[23] A. Liutkus and P. Leveau, "Separation of music+effects sound track from several international versions of the same movie," in *Proc. 128th Audio Engineering Society (AES) Convention*, London, United Kingdom, May 2010.

[24] T. Gerber, M. Dutasta, L. Girin, and C. Févotte, "Professionally-produced music separation guided by covers," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, Oct. 2012, pp. 85–90.

[25] R. Hennequin, J. J. Burred, S. Maller, and P. Leveau, "Speech-guided source separation using a pitch-adaptive guide signal model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italie, May 2014, pp. 6672–6676.

[26] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Audio source separation using multiple deformed references," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisboa, Portugal, Sept. 2014, pp. 311–315.

[27] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-Informed Audio Source Separation. Example-Based Approach Using Non-Negative Matrix Partial Co-Factorization," *Journal of Signal Processing Systems*, pp. 1–15, 2014.

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

[29] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[30] 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, 2007.

[31] L. Catanese, N. Souviraà-Labastie, B. Qu, S. Campion, G. Gravier, E. Vincent, and F. Bimbot, "MODIS: an audio motif discovery software," in *Show & Tell - Interspeech 2013*, Lyon, France, 2013.

[32] A. Muscariello, G. Gravier, and F. Bimbot, "Unsupervised motif acquisition in speech via seeded discovery and template matching combination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2031–2044, Sept. 2012.

[33] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraà-Labastie, and F. Bimbot, "BL-Database: A french audiovisual database for speech driven lip animation systems," INRIA, Technical report RR-7711, 2011.

[34] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[35] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[36] D. P. W. Ellis. (2003) Dynamic time warping in Matlab. [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/

[37] ——. (2005) PLP and RASTA (and MFCC, and inversion) in Matlab. [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/

Nathan Souviraà-Labastie graduated in Electronics, Computer Engineering and Embedded systems from Institut National des Sciences Appliquées (INSA Rennes) in 2011. He is now a PhD candidate at the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), Rennes, France. His research interests are in the fields of audio source separation, speech recognition, and audio indexing.

Anaik Olivero received the M.Sc. degree in acoustics from the Université de Provence in 2008, and the Ph.D. degree from Aix-Marseille University in 2012 after completing a thesis on audio signal processing. Currently, she is a post-doctoral researcher at Institut National de Recherche en Informatique et Automatique (Inria) in Rennes. Her research interests are in signal processing and machine learning applied to audio signals.

Emmanuel Vincent is a research scientist with Inria (Nancy, France). He received the Ph.D. degree in music signal processing from the Institut de Recherche et Coordination Acoustique/Musique (Paris, France) in 2004 and worked as a research assistant with the Centre for Digital Music at Queen Mary, University of London (United Kingdom), 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 a founder of the series of Signal Separation Evaluation Campaigns and CHiME Speech Separation and Recognition Challenges. He was an associate editor for IEEE Transactions on Audio, Speech, and Language Processing. He is a Senior Member of the IEEE.

Frédéric Bimbot graduated as a telecommunication engineer from ENST, Paris, France, in 1985, received the B.A. degree in linguistics from Sorbonne Nouvelle University, Paris III, and the Ph.D. degree in signal processing (speech synthesis using temporal decomposition) from ENST in 1988. In 1990, he joined CNRS as a permanent researcher, worked with ENST for seven years, and then moved to IRISA (CNRS and Inria), Rennes, France. He is now a senior researcher with CNRS. His research is focused on speech and audio analysis, speaker recognition, music content modeling, and audio source separation. He is in charge of the coordination of the D5 Department (digital signals and images, robotics) at IRISA.