Carabias-Orti *et al. EURASIP Journal on Advances in Signal Processing* 2013, **2013**:184
http://asp.eurasipjournals.com/content/2013/1/184

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings

Julio J Carabias-Orti[1*], Máximo Cobos[2], Pedro Vera-Candeas[3] and Francisco J Rodríguez-Serrano[3]

Abstract

Close-microphone techniques are extensively employed in many live music recordings, allowing for interference rejection and reducing the amount of reverberation in the resulting instrument tracks. However, despite the use of directional microphones, the recorded tracks are not completely free from source interference, a problem which is commonly known as microphone leakage. While source separation methods are potentially a solution to this problem, few approaches take into account the huge amount of prior information available in this scenario. In fact, besides the special properties of close-microphone tracks, the knowledge on the number and type of instruments making up the mixture can also be successfully exploited for improved separation performance. In this paper, a nonnegative matrix factorization (NMF) method making use of all the above information is proposed. To this end, a set of instrument models are learnt from a training database and incorporated into a multichannel extension of the NMF algorithm. Several options to initialize the algorithm are suggested, exploring their performance in multiple music tracks and comparing the results to other state-of-the-art approaches.

1 Introduction

Multitrack audio recording techniques are based on capturing and recording individual sound sources into multiple discrete audio channels. Once all the sound sources have been recorded, the individual tracks are processed and mixed down to a number of mixture channels that depends on the specific audio reproduction format. Multitrack recording techniques can be broadly classified into live recording and track-by-track recording techniques. In the latter type, the performers are individually recorded one after another, resulting in almost perfectly isolated instrument tracks. On the other hand, in live audio recordings, the source signals, which share the acoustic space, are all acquired simultaneously during the performance [1]. This leads to the well-known microphone leakage problem: the sounds coming from the concurrent sources are picked up by microphones others than the ones intended for the specific sources [2]. To address this issue in close-miking techniques, a directional microphone is placed relatively close to an instrument, reducing the interference from other sources and the effect of room reverberation. Other mechanical and signal processing devices, such as absorbing barriers or noise gates, are also employed by sound engineers to mitigate this problem, but they only solve the problem partially, being most effective when used with transient signals [3].

Sound source separation (SSS) techniques have been suggested as a potential solution for the microphone leakage problem in multitrack live recordings [3,4]. In general, the aim of SSS is to recover each source signal from a set of audio mixtures. SSS techniques can be broadly divided into blind source separation (BSS) and informed source separation (ISS) algorithms. BSS methods are especially popular in the statistical signal processing and machine learning areas, where the term blind emphasizes that very little information about the sources or the mixing process is known a priori [5]. Techniques such as principal

*Correspondence: julio.carabias@upf.edu
[1]Music Technology Group (MTG), Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona 08018, Spain
Full list of author information is available at the end of the article

component analysis (PCA), independent component analysis (ICA) or nonnegative matrix factorization (NMF) [6] have been introduced both to reduce the dimensionality and to explain the whole data by a few meaningful elementary objects. In fact, many BSS approaches are closely related to ICA, where the sources are assumed to be statistically independent and non-Gaussian. Most of these approaches are oriented to the determined separation problem, i.e., the number of sources equals the number of mixture signals. When the number of sources is greater than the number of mixtures, the problem is said to be underdetermined, and the underlying assumptions usually involve the sparsity of the sources under some suitable representation such as the time-frequency domain [7,8]. Moreover, the assumptions may also differ depending on the acoustic environment, leading to instantaneous or convolutive separation methods. Instantaneous mixing models (IMM) assume a mixing matrix made up of scalar coefficients while convolutive models are often based on the estimation of unmixing filters [9]. When working in the frequency domain, the mixture can be assumed to be instantaneous at each frequency bin and standard approaches such as ICA can be applied by following a subband approach [10]. However, due to the ICA permutation ambiguity, an alignment procedure requiring some additional information is necessary to group the resulting components into estimated source signals. When the signal model is assumed to be nonnegative, NMF provides a meaningful structure of the audio data, which in this case is obtained from the magnitude or energy spectrograms. NMF methods have been shown to be specially useful for musical analysis tasks [11], including not only SSS but also others, such as automatic music transcription [12] or acoustic space characterization [13]. NMF is based on decomposing the spectrogram audio data into a sum of elementary spectral patterns with time-varying gains. While NMF was originally proposed in the context of monaural SSS, other extensions have been developed for dealing with multichannel audio mixtures [14]. As a result, NMF approaches are progressively becoming a promising solution to multichannel SSS. However, spectral patterns learnt by NMF-based approaches are often hard to interpret and lack explicit semantics. To overcome this issue, many algorithms constrain the original NMF to obtain musically meaningful patterns, for example, by considering a parametric model. In this context, the spectral patterns can be described by harmonic combs [15-17], spectrally and/or temporally localized Gaussians [18,19], or by using a source/filter model [20-22].

In contrast to BSS, ISS methods depart from an available prior information, which can be under the form of specific information about the sources, the mixing process, or additional modalities [23]. For example, an ISS method which is oriented to close-miking live music recordings could exploit the properties of this specific setup: each microphone signal contains one of the sources significantly enhanced over the others due to both the directional properties of the sensors and to their placement. In this context, Kokkinis and Morjopoulos [3] showed that under a close-miking assumption, a relatively simple Wiener filter outperforms some convolutive BSS algorithms. However, more sophisticated methods making use of additional prior information can be developed by considering a supervised separation framework. For example, musical score information can be used if the score and audio are well aligned [24-27]. Spectral information can be considered by using instrument models when the instruments are known in advance [28,29]. Other kinds of information, such as high-level musicological knowledge, have been recently introduced by Fuentes et al. [30], using recent advances in shift-invariant analysis of musical data. Regarding factorization methods, an important issue to take into account is the initialization/constrain of the parameters. In this context, Hurmalainen et al. [31] proposed a method for automatic adaptation of learnt clean speech source models to deal with noise in a speech separation and recognition task. Furthermore, Fitzgerald [32] presented a framework that allows the user to interact with the tensor factorization method to improve the performance in an adaptive way. Finally, the prior information can be the sources themselves. This knowledge enables the computation of side information, which is small enough to be inaudibly embedded into the mixtures. At a decoding step, this small side information is used along with the mixtures to recover the sources. Following this scheme, Liutkus et al. [33] proposed a system coding approach that permits very reliable transmission of the sources with a small amount of side information.

In this paper, an informed NMF-based SSS method is presented to tackle the microphone leakage problem in multichannel close-microphone recordings. To this end, several assumptions are taken on the mixing environment, affecting problem dimensionality, direct-to-reverberant sound ratio and available instrument priors. In this context, it is assumed that the number of source signals is equal or less than the number of microphone signals, having each mixture signal a predominant direct-sound source resulting from a close-miking recording setup. Therefore, since the predominant source is captured with a high direct-to-reverberant ratio, a instantaneous model can be reasonably assumed, significantly simplifying the separation task. Since the method is constrained to be nonnegative, panning matrix is used to determine the mixing process. Moreover, instrument model priors are obtained by means of a learning stage using a training database. The usefulness of these models is twofold. On the one hand, they enable an accurate

estimation of the panning matrix. On the other hand, they simplify the separation stage by reducing the factorization to the estimation of instrument time-varying gains.

The paper is structured as follows. Section 2 provides an overview of the proposed SSS system and describes the fundamentals of NMF-based separation and instrument modeling. Section 3 describes the proposed multichannel extension for informed NMF-based separation using learnt instrument models. Panning matrix estimation and NMF-based separation are described in detail, explaining how the output of an automatic music transcription stage is used to discriminate single-source time-frequency zones. Section 4 describes the experiments conducted by using several music pieces in a simulated close-microphone setup and evaluates the separation performance by using objective measures. Finally, Section 5 summarizes the conclusions of this work.

2 Model description and background

2.1 System overview

Figure 1 shows a diagram of the proposed NMF-based separation system. Note that the system is composed of two main blocks: a panning matrix estimation block and the actual source separation block. Both blocks need as an input the spectrograms of the mixture microphone signals and a set of instrument basis functions calculated from an available training database. The panning matrix estimation procedure is based on the discrimination of time-frequency zones with minimum overlap between the concurrent instruments, which is performed by using the output of an automatic transcription stage. The estimated panning matrix is then fed to the NMF-based separation stage, which also uses the modeled instrument basis to estimate the magnitude spectrograms of the original sources. These spectrograms are finally used to recover the actual sources by constructing a Wiener mask that is applied over the input spectrograms. The symbols appearing in this diagram are defined throughout the description of the proposed approach in the following sections.

2.2 NMF background

Factorization-based approaches have been intensively applied in signal processing applications, including single-channel audio source separation [29,34-36]. The principle is that an audio spectrogram can be decomposed as a linear combination of spectral basis functions. In such a model, the short-term magnitude (or power) spectrum of the signal $x(f,t)$ in time-frame t and frequency f is modeled as a weighted sum of basis functions as

$$x(f,t) \approx \hat{x}(f,t) = \sum_{n=1}^{N} b_n(f) g_n(t), \tag{1}$$

where $g_n(t)$ is the gain of the basis function n at frame t, and $b_n(f)$, $n = 1,\ldots,N$ are the bases. Note that this approach holds under two different configurations:

(a) Strong sparsity (only one source active per TF bin).
(b) Under a local stationarity assumption. In that case, model in Equation 1 does not hold for magnitude spectrograms, but rather holds in average for power spectrograms [37-39].

Therefore, whenever model in Equation 1 is chosen, either assumptions (a) or (b) are supposed to hold, and the time-frequency (TF) representation considered is either magnitude or power spectrogram for (a) or power spectrogram only for (b).

When dealing with musical instrument sounds, ideally, each basis function can represent a single pitch, and the corresponding gains contain information about the

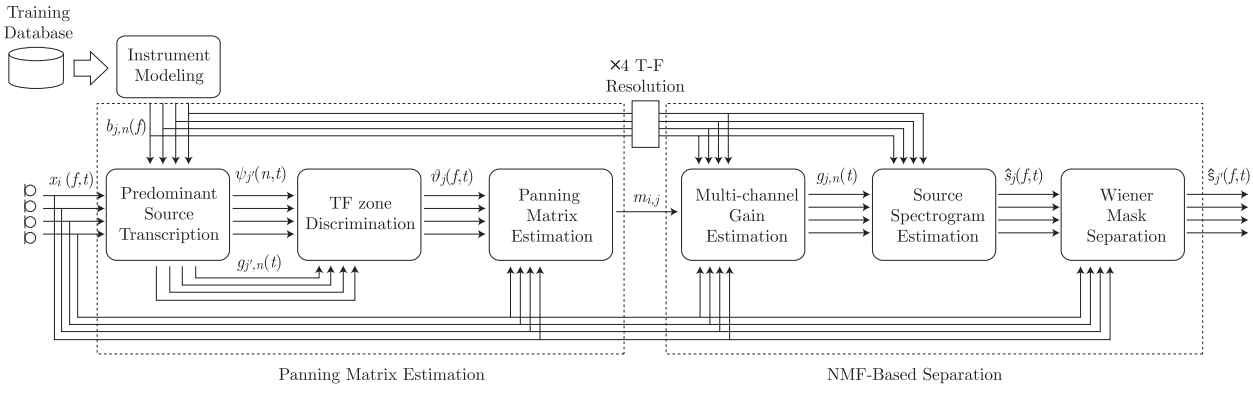


Figure 1 System diagram.

onset and offset times of notes having that pitch. Several works [16,17,22,40-43] proposed to restrict the model in Equation 1 to be harmonic. The harmonicity constraint is particularly useful for the analysis and separation of musical audio signals since, by using this constraint, each basis can define a single fundamental frequency. In previous works, the authors have introduced this constraint in the model presented in Equation 1 by requiring that a distinct basis function represents each note of each instrument:

$$b_{j,n}(f) = \sum_{h=1}^{H} a_{j,n}(h) G\left(f - hf_0(n)\right), \tag{2}$$

where $b_{j,n}(f)$ are the bases for each note n of instrument j; $n = 1,\ldots,N$ is defined as the pitch range for instrument $j = 1,\ldots,J$, where J is the total number of instruments present in the mixture; $h = 1,\ldots,H$ is the number of harmonics; $a_{j,n}(h)$ is the amplitude of harmonic h for note n and instrument j; $f_0(n)$ is the fundamental frequency of note n; $G(f)$ is the magnitude spectrum of the window function; and the spectrum of a harmonic component at frequency $hf_0(n)$ is approximated by $G(f - hf_0(n))$. The model for the magnitude spectra of a music signal is then expressed as

$$\hat{x}(f,t) = \sum_{j=1}^{J}\sum_{n=1}^{N}\sum_{h=1}^{H} g_{j,n}(t) a_{j,n}(h) G(f - hf_0(n)), \tag{3}$$

where the time gains $g_{j,n}(t)$ and the harmonic amplitudes $a_{j,n}(h)$ are the parameters to be estimated.

2.3 Augmented NMF for parameter estimation

The nonnegativity of the parameters is widely used in music transcription [12,16,17,22] and source separation [36,44]. Under the nonnegativity restriction, the factorization parameters of Equation 3 can be estimated by minimizing the reconstruction error between the observed $x(f,t)$ and the modeled $\hat{x}(f,t)$ spectrograms. In several recent works [16,45,46], the cost function to be minimized is the beta-divergence:

$$D_\beta(x(f,t)|\hat{x}(f,t)) = \begin{cases} \sum_{f,t} \frac{1}{\beta(\beta-1)}\left(x(f,t)^\beta + (\beta-1)\hat{x}(f,t)^\beta - \beta x(f,t)\hat{x}(f,t)^{\beta-1}\right) & \beta \in (0,1)\cup(1,2] \\ \sum_{f,t} x(f,t)\log\frac{x(f,t)}{\hat{x}(f,t)} - x(f,t) + \hat{x}(f,t) & \beta = 1 \\ \sum_{f,t} \frac{x(f,t)}{\hat{x}(f,t)} + \log\frac{x(f,t)}{\hat{x}(f,t)} - 1 & \beta = 0 \end{cases} \tag{4}$$

The beta-divergence includes in its definition the most popular cost functions. When $\beta = 2$, the beta-divergence is equivalent to the Euclidean (EUC) distance. The Kullback-Leibler (KL) divergence is obtained when $\beta = 1$, and the Itakura-Saito (IS) divergence is computed when $\beta = 0$. To estimate the parameters in Equation 1 that minimize the cost function, Lee et al. [47] proposed an iterative algorithm based on multiplicative update (MU) rules. Under these rules, $D(x(f,t)|\hat{x}(f,t))$ is shown to be nonincreasing at each iteration while ensuring the nonnegativity of the bases and the gains. These MU rules are obtained applying diagonal rescaling to the step size of the gradient descent algorithm (see [47] for further details). The MU rule for each scalar parameter θ_l is obtained as follows. First, the derivative $\nabla_{\theta_l} D$ of the cost function with respect to θ_l is expressed as a difference between two positive terms $\nabla^+_{\theta_l} D$ and $\nabla^-_{\theta_l} D$ [46]. Then, θ_l is updated by

$$\theta_l \leftarrow \theta_l \frac{\nabla^-_{\theta_l} D(x(f,t)|\hat{x}(f,t))}{\nabla^+_{\theta_l} D(x(f,t)|\hat{x}(f,t))}. \tag{5}$$

The main advantage of the MU in Equation 5 is that it ensures nonnegativity of all parameters provided that they are nonnegative at initialization.

2.4 Instrument modeling

As demonstrated in [22], when appropriate training data are available, it is advantageous to learn the instrument-dependent bases in advance and fix them during the analysis of the signals. In fact, this approach has been shown to perform well when the conditions of the music scene do not differ too much between the training and the test data. Here, we have used an approach similar to [43]. Specifically, the amplitudes of each note of a musical instrument $a_{j,n}(h)$ are learnt in advance by using the Real World Computing (RWC) music database [48,49] as a training database of solo instruments playing isolated notes (more details on the Section 4.2.). Let the ground-truth transcription of the training data be represented by $R_{j,n}(t)$ as a binary time/frequency matrix for each j instrument. The frequency dimension represents the musical instrument digital interface (MIDI) scale and the time dimension t represents the frames. At the training

stage, gains are initialized with $R_{j,n}(t)$, which is known in advance for the training database. Thus, gains are set to unity for each pitch at those time frames where the instrument is active while the rest of the gains are set to zero. Note that gains initialized to zero remain at zero because of the multiplicative update rules, and therefore the frame is represented only with the correct pitch.

The MU rules that minimize the beta-divergence for the parameters of the model are computed from Equation 5 for the amplitude coefficients and the gains as follows:

$$a_{j,n}(h) \leftarrow a_{j,n}(h)\frac{\sum_{f,t} x(f,t)\hat{x}(f,t)^{\beta-2} g_{j,n}(t) G(f-hf_0(n))}{\sum_{f,t} \hat{x}(f,t)^{\beta-1} g_{j,n}(t) G(f-hf_0(n))}. \tag{6}$$

$$g_{j,n}(t) \leftarrow g_{j,n}(t)\frac{\sum_{f,m} x(f,t)\hat{x}(f,t)^{\beta-2} a_{j,n}(h) G(f-hf_0(n))}{\sum_{f,m} \hat{x}(f,t)^{\beta-1} a_{j,n}(h) G(f-hf_0(n))}. \tag{7}$$

The training procedure is summarized in Algorithm 1.

Algorithm 1 Instrument modeling algorithm

1: Compute $x(f,t)$ from a solo performance for each instrument in the training database
2: Initialize gains $g_{j,n}(t)$ with the ground-truth transcription $R_{j,n}(t)$ and $a_{j,n}(h)$ with random positive values.
3: Update the gains using Equation 6.
4: Update the bases using Equation 7.
5: Repeat steps 2 to 3 until the algorithm converges (or maximum number of iterations is reached).
6: Compute basis functions $b_{j,n}(f)$ for each instrument j using Equation 2.

The training algorithm computes the basis functions $b_{j,n}(f)$ required at the factorization stage for each instrument. These instrument-dependent basis functions $b_{j,n}(f)$ are known and held fixed, therefore, the factorization of new signals of the same instrument can be reduced to the estimation of the gains $g_{j,n}(t)$.

3 Proposed extension to multichannel

The previously described NMF-based model is suitable for single-channel data. However, most music recordings are available in a multichannel format, being stereo the most common. To deal with multichannel audio data, an extension of the standard NMF model is required. In the literature, multichannel extensions of NMF have already been considered, either by stacking up the spectrograms of each channel into a single matrix [50] or by equivalently considering nonnegative tensor factorization (NTF) under a parallel factor analysis (PARAFAC) structure, where the channel spectrograms form the slices of a 3-valence tensor [42,51,52].

In this paper, we propose an extended multichannel NMF model that is specifically designed for close-microphone music recordings. While this kind of recordings are not usually commercially distributed, many of the raw recordings used in the studio during the mixing process share many similarities among them. The particularities of this scenario define a set of assumptions that are considered in the proposed NMF algorithm:

- Problem dimensionality: The proposed method is designed for an over-determined scheme, that is, $I \geq J$ where I is the number of channels and J the number of sources.
- Single predominant source: For each channel i, there is a single predominant source j' that corresponds to a music instrument which is known in advance.
- Mixing model: In this work, instantaneous mixing of point sources is considered. Note that the actual mixing process in a close-microphone recording is convolutive. However, since the predominant source is captured with a high direct-to-reverberant ratio, a instantaneous model can be reasonably assumed to simplify the processing. Still, the proposed method can readily be extended to the case of a convolutive mixture, simply by assuming a mixing matrix that varies over frequency [14].
- Input representation: More details about the TF representation are given in the experimental section.

Now, under the previously detailed assumptions, let $\underline{\mathrm{x}}_i(f,t)$ be the complex-valued short-time Fourier transform (STFT) of channel $i = 1,\ldots,I$ and I the number of channels. Then, the complex-valued STFT $\underline{\mathrm{x}}_i(f,t)$ for each channel of the multichannel data can be expressed as

$$\underline{\mathrm{x}}_i(f,t) \approx \hat{\underline{\mathrm{x}}}_i(f,t) = \sum_{j=1}^{J} m_{i,j}\underline{\mathrm{s}}_j(f,t), \tag{8}$$

where $\hat{\underline{\mathrm{x}}}_i(f,t)$ is the estimation of the complex-valued STFT for each i channel; $\underline{\mathrm{s}}_j(f,t)$ is the estimation of the complex-valued STFT generated by the source $j = 1,\ldots,J$; J is the number of sources; and the scalar coefficients $m_{i,j}$ define a $I \times J$ panning matrix $\mathbf{M}$ that measures the multichannel contribution of source j to the data. Note that the mixing coefficients are defined in function of the kind of spectrogram used (i.e., magnitude or power spectrogram). On the one hand, if we are considering magnitude spectrograms, mixing coefficients can be defined as $|m_{i,j}|$, usually called the mixing matrix. On the other hand, for power spectrograms mixing coefficients are defined as $|m_{i,j}|^2$.

The studied multichannel sound source separation problem is illustrated in Figure 2. Assuming that each source is generated by a single musical instrument, we can use the model described in Section 2, that is, the spectrogram of each source j can be modeled as a product of two nonnegative matrices $g_{j,n}(t)$ and $b_{j,n}(f)$, such that

$$s_j(f,t) \approx g_{j,n}(t) b_{j,n}(f), \tag{9}$$

where $s_j(f,t)$ is the estimation of the spectrogram generated by source j. Considering the harmonicity constraint imposed in Equation 2, we can redefine the source spectrogram as

$$s_j(f,t) \approx \sum_{n=1}^{N} g_{j,n}(t) \sum_{h=1}^{H} a_{j,n}(h) G(f - hf_0(n)). \tag{10}$$

Finally, the model for the magnitude spectrogram of a multichannel music signal is then obtained as

$$\hat{x}_i(f,t) = \sum_{j=1}^{J} |m_{i,j}| \sum_{n=1}^{N} g_{j,n}(t) \sum_{h=1}^{H} a_{j,n}(h) G(f - hf_0(n)). \tag{11}$$

3.1 Panning matrix estimation

The estimation of the panning matrix is performed in two steps. First, an NMF-based automatic transcription method is applied in order to estimate the active notes of the predominant source at each channel. Then, the estimated transcription of the predominant source for each channel is used to discriminate those TF zones in which the sources are presented in an isolated way. Finally, the panning matrix is computed using this information.

3.1.1 Predominant source transcription method

In this step, we describe two NMF-based methods, one for monophonic and the other for polyphonic signals, to estimate the transcription of the predominant source for each channel individually. These methods were previously developed by the authors in [43] in the context of monaural mixtures. The methods are supervised, requiring fixed basis functions trained using the instrument modeling procedure in Section 2.4. The aim here is to estimate the transcription of the predominant source j' at each channel i. This information must be known in advance in order to define the proper basis functions $b_{j',n}(f)$.

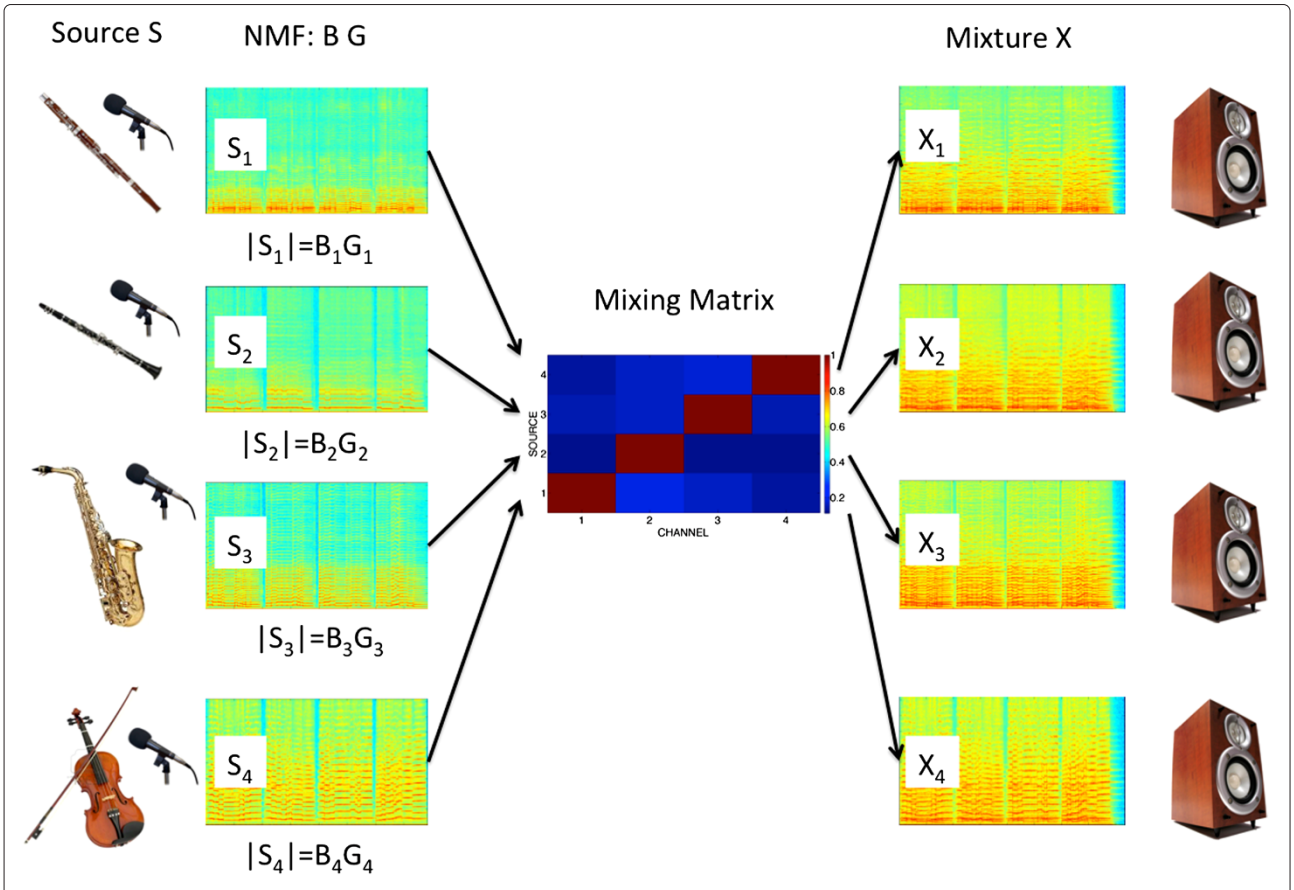


Figure 2 Representation of instantaneous mixing system and formulation of multichannel NMF problem.

The supervised NMF methods with learnt instrument basis functions are described below:

1. *Monophonic sources* In the case of monophonic sources, we propose to use the real-time single-pitch constrained method proposed in [43]. In this transcription method, the optimum combination of notes $n_{\text{opt}}(i,t)$ is chosen to minimize the beta-divergence function at channel i and frame t under the assumption that only one gain $g_{j',n}(t)$ is nonzero at each frame for channel i, being j' the predominant source. Assuming a single predominant source j' for each channel i, the signal model with the single-combination constraint can be defined as follows:

$$x_{i,t}(f) \approx s_{j',n_{\text{opt}},t}(f) = g_{j',n_{\text{opt}},t} b_{j',n_{\text{opt}}}(f), \tag{12}$$

where $n_{\text{opt}}(i,t)$ is defined for each channel i as

$$n_{\text{opt}}(i,t) = \arg\min_{n=1,\ldots,N} D_\beta\left(x_{i,t}(f) | g_{j',n,t} b_{j',n}(f)\right), \tag{13}$$

that is, the spectrum for each channel i at each frame t is approximated by the projection of the predominant source j' for the optimum note n_{opt} at frame t. As an advantage, the model of Equation 12 allows the gains to be computed directly from the data and the trained amplitudes without the need for an iterative algorithm.
The beta-divergence at note n and frame t for the predominant source j' at channel i is obtained as

$$\begin{aligned} D_\beta(x_{i,t}(f) | g_{j',n,t} b_{j',n}(f)) = \sum_f \frac{1}{\beta(\beta-1)} \big(& x_{i,t}(f)^\beta \\ & + (\beta-1)(g_{j',n,t} b_{j',n}(f))^\beta \\ & - \beta x_{i,t}(f)(g_{j',n,t} b_{j',n}(f))^{\beta-1} \big). \end{aligned} \tag{14}$$

The value of the gain for channel i, source j', note n and frame t is obtained by minimizing Equation 14. This minimization has a direct solution, since the value of the gain for note n and frame t is a scalar:

$$g_{j',n,t} = \frac{\sum_f x_{i,t}(f) b_{j',n}(f)^{(\beta-1)}}{\sum_f b_{j',n}(f)^\beta}. \tag{15}$$

Finally, the optimum note at each frame for each channel is selected as the note that minimizes the beta-divergence at each frame and channel:

$$n_{\text{opt}}(i,t) = \arg\min_{n=1,\ldots,N} D_\beta\left(x_{i,t}(f) | \frac{\sum_f x_{i,t}(f) b_{j',n}(f)^{(\beta-1)}}{\sum_f b_{j',n}(f)^\beta} b_{j',n}(f)\right), \tag{16}$$

where the proposed solution is valid for $\beta \in [0,2]$. To summarize, the monophonic predominant source transcription (MPST) method is detailed in Algorithm 2.

Algorithm 2 Monophonic predominant source transcription (MPST) method

```
for i = 1 to I do
  Initialize b_{j',n}(f) with the values learnt in Section 2.4
  and g_{j',n}(t) with random positive values.
  for t = 1 to T do
    for n = 1 to N do
      Compute the gains g_{j',n,t} using Equation 15.
    end for
    Choose the note n_opt(i,t) using Equation 16
  end for
end for
```

2. *Polyphonic sources* In the case of polyphonic sources, the method presented in [43] is used to obtain the transcription of the predominant source j' at each channel i. However, we highlight the fact that any polyphonic estimation procedure may be used at this stage, e.g., the one presented in [30]. In this paper, the applied method is analogous to the classical Euclidean-NMF using the gradient descent algorithm, but no iterative process is required, allowing its use for real-time problems. In the case of $\beta = 2$ (Euclidean distance), Equation 4 for each channel i can be expressed in matrix notation as

$$D_{\beta,i} = \left\| \mathbf{X}_i - \mathbf{B}_{j'} \cdot \mathbf{G}_{j'} \right\|_2, \tag{17}$$

where $\mathbf{X}_i$ is the signal input matrix at channel i in time and frequency, j' is the predominant source at channel i, $\mathbf{B}_{j'}$ the j' source basis matrix, and $\mathbf{G}_{j'}$ the j' source gain matrix. Then we can examine this factorization as a reduced-rank basis decomposition so that $\mathbf{X}_i = \mathbf{B}_{j'} \cdot \mathbf{G}_{j'}$ and, subsequently, the gains can be estimated in just one step:

$$\mathbf{G}_{j'} = \mathbf{A}_{j'} \cdot \mathbf{X}_i, \tag{18}$$

where $\mathbf{A}_{j'} \in \mathbb{R}^{\geq 0, N \times F} = \mathbf{B}_{j'}^\dagger$ and the † operator is the Moore-Penrose matrix inverse.
As commented in [43], the transcription results with $\beta = 2$ are poor in comparison with other values of β. Therefore, to improve the performance of the method, a candidate selection stage using the

previously explained MPST method will be applied. As a result, the method for polyphonic sources is restricted to have only a few gains that are not zero for each instrument at each frame. Detailed information about the method can be found in [43]. Algorithm 3 describes the computational procedure in the proposed transcription method for polyphonic sources.

Algorithm 3 Polyphonic predominant source transcription (PPST) method

for $i = 1$ to I **do**
  Initialize $b_{n,j'}(f)$ to the trained instrument models (Section 2.4)
  for $t = 1$ to number of frames **do**
    Compute the MPST for channel i with Equation 15 and Equation 14
    Select the C notes that causes the lowest Beta-divergence at frame t, C being the number of note candidates
  end for
  Generate basis dictionary $\mathbf{B_{j'}}$ with the candidate basis functions $b_{n_c,j'}(f)$.
  Estimate the candidate gains using Equation 18. The rest of values are set to 0.
end for

Finally, given the time-varying amplitudes of each predominant source $g_{j',n}(t)$, the transcription matrix for the predominant source j' at channel i is computed using the method proposed in [16,17,53]. Specifically, we determine whether a note is active or not on a frame-by-frame basis according to the following equation:

$$\psi_{j'}(n,t) = g_{j',n}(t) \geq \left(10^{T/20} \max_{nt} g_{j',n}(t)\right) \tag{19}$$

where $\psi_{j'}(n,t)$ is the resulting transcription composed of binary values and T is a fixed detection threshold in decibels (dB) that can be either set manually or learnt from training data. Note that in the case of MPST, where only one note is active at a time, a threshold is used to discard those notes actives during silence intervals.

3.1.2 Time-frequency zones discrimination

Once the transcription procedure of the predominant source at each channel i has been performed, time-frequency zones in the mixture corresponding only to predominant sources must be discriminated in order to estimate the panning matrix. These zones are assumed to be free from source overlapping, thus, partials from the predominant source are likely not to be corrupted by notes from the rest of instruments. In fact, the information resulting from overlapped partials at each frame is considered as corrupted and it is not used to estimate the panning matrix.

Figure 3 shows some overlapping partials from two notes. The spectrum of a bassoon note (56, blue) and saxophone note (65, red) are drawn. It can be observed how the fifth partial of the bassoon note and the third partial from the saxophone note are overlapped. The same occurs with the tenth partial of the bassoon note and the sixth of the saxophone. These overlapping points may be avoided at the panning matrix estimation process. Therefore, the transcription $\psi_{j'}(n,t)$ of the predominant source at each channel is used for determining those time-frequency zones in which each instrument j is not free from interferences of the rest of instruments. These time-frequency zones are here defined as an overlapping mask $\vartheta_j(f,t)$. It must be stressed that this approach requires to use at least one channel per instrument, so the number of channels must be equal or higher than the number of sources, $I \geq J$, i.e., the problem must be determined or over-determined. To estimate the overlapping mask $\vartheta_j(f,t)$, an energy estimation per source is computed using Equation 9 with the initialization from the estimated predominant source transcription for the gains $g_{j',n}(t)$ and the basis functions fixed from the instrument modeling stage (Section 2.4). If the estimated energy for one source is denoted as $s_{1,t}(f)$ and the other as $s_{2,t}(f)$, the frequencies where $\frac{s_{1,t}(f)}{s_{2,t}(f)} \geq 0.1$ are selected as overlapped ones and $\vartheta_j(f,t)$ is conveniently updated by setting to zero those overlapped frequencies f at the corresponding frame t. This overlap estimation is performed over every pair of sources.

An example of this stage is illustrated in Figure 4. It can be seen how the predominant source transcription from each channel spectrogram (Figure 4a,d,g,j leads to Figure 4b,e,h,k). Finally, an overlapping mask $\vartheta_j(f,t)$ is defined for each source by using the explained energy-based method with the fixed basis functions and the estimated predominant source transcription. The resulting overlapping masks are displayed in Figure 4c,f,i,l for each predominant source.

3.1.3 Energy-based panning matrix estimation

Once the overlapping mask $\vartheta_j(f,t)$ is computed for all the channels, the panning matrix is estimated (see Figure 1). The proposed method computes each panning coefficient as the relation between the norm of each instrument at each channel in its time-frequency region and the norm of each instrument at its predominant channel in its time-frequency region (the panning coefficient for the instrument at its predominant channel is supposed to be 1). Note that the panning estimation method does not include the phase of the coefficient; in other words, the proposed method estimates the matrix coefficients

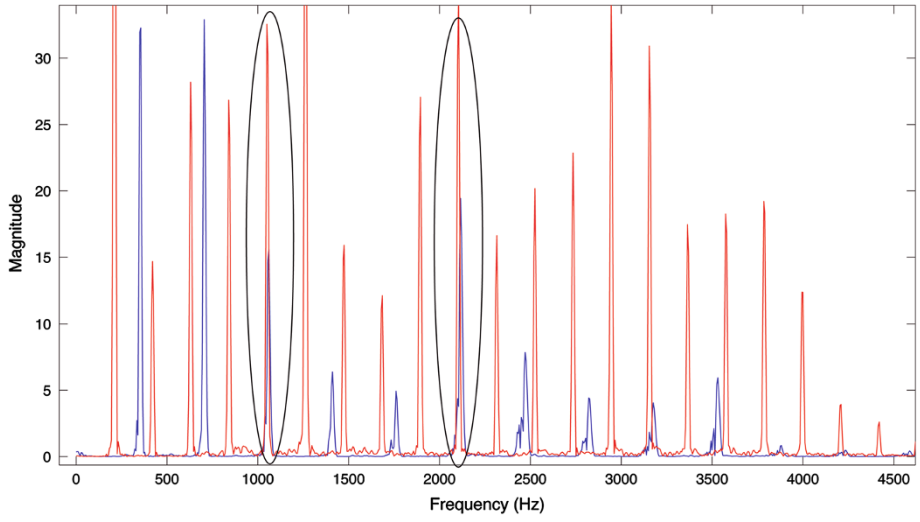


Figure 3 Spectral representation of a note of clarinet (65, blue) and bassoon (56, red) with almost two overlapped partials.

for magnitude spectrograms. The energy-based panning matrix estimation is detailed in Algorithm 4, where j' is the predominant source for channel i, $\circ$ is the Hadamard product and $\|\cdot\|_2$ is the 2-norm (Euclidean distance).

Algorithm 4 Panning matrix estimation method

for $i = 1$ to I **do**
  j' is the predominant source for channel i.
  for $j = 1$ to J **do**
    if $j == j'$ **then**
      $m(i,j) = 1$
    else
      $m(i,j) = \|x_i(f,t) \circ \vartheta_j(f,t)\|_2 \,/\, \|x_i(f,t) \circ \vartheta_{j'}(f,t)\|_2$
    end if
  end for
end for

Therefore, Algorithm 4 computes the panning matrix as the quotient between the contribution of each source to the channel spectrogram against the contribution of the predominant source.

3.2 Multichannel SSS

In this stage, we will estimate the reconstruction of each source using the fixed basis functions $b_{j,n}(f)$ from the instrument modeling stage (Section 2.4) and the previously estimated panning matrix $\mathbf{M}$ with elements $m_{i,j}$. First of all, we will define the update rule in Equation 7 for the gains as follows:

$$g_{j,n}(t) \leftarrow g_{j,n}(t) \frac{\sum_{f,i} m_{i,j} b_{j,n}(f) x_i(f,t) \hat{x}_i(f,t)^{\beta-2}}{\sum_{f,i} m_{i,j} b_{j,n}(f) \hat{x}_i(f,t)^{\beta-1}}. \quad (20)$$

Then, classical augmented NMF factorization with MU rules is applied to estimate the gains corresponding to each source j in the multichannel mixture. The process is detailed in Algorithm 5.

Algorithm 5 Multichannel signal gain estimation method

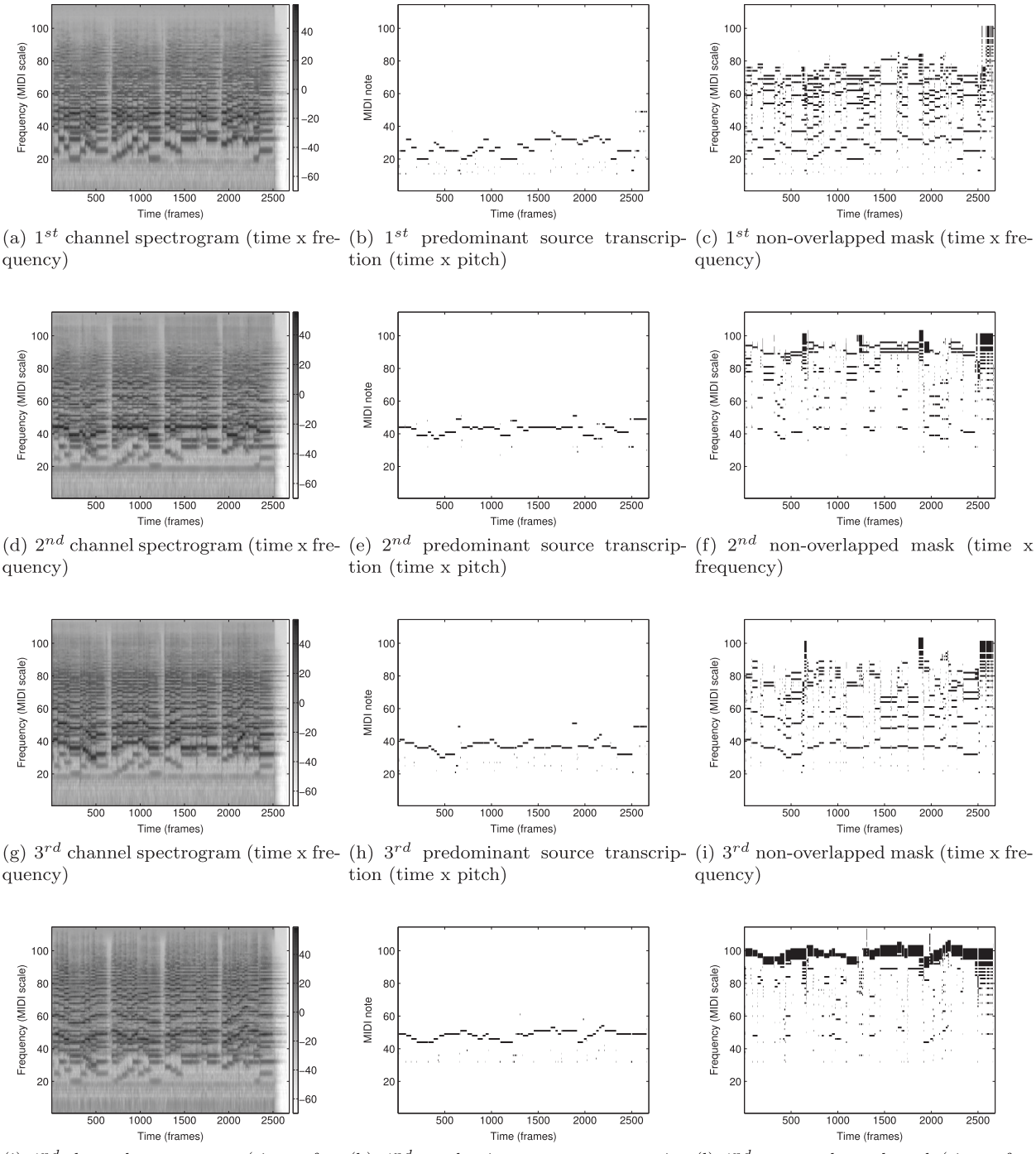
1. Initialize $b_{j,n}(f)$ and $m_{i,j}$ with the values learned in Section 2.3 and Section 3.1, respectively. Use random positive values to initialize $g_{j,n}(t)$.
2. Update the gains using Equation 20.
3. Repeat step 2 until the algorithm converges (or maximum number of iterations is reached).

3.2.1 *Ideal Wiener masks*

The source separation consists of estimating the complex amplitude at each time-frequency cell for each source. Some systems use binary separation, which means that the entire energy of a bin is assigned to a single source. However, it has been demonstrated that better results can be obtained with a nonbinary decision, i.e., distributing the energy proportionately over all the sources. The use of separation Wiener masks is common in the source separation literature [37-39]. The Wiener filter method for instantaneous mixing models is described below.

Let $\underline{s}_j(f,t)$ be the complex-valued STFT for source j at TF bin (f,t). If we assume that source j is locally stationary, it is equivalent to assume that all the T elements of its STFT are independent and that each element $\underline{s}_j(f,t)$ is distributed with respect to a complex-centered Gaussian distribution:

$$\underline{s}_j(f,t) \sim \mathcal{N}(0, |\underline{s}_j(f,t)|^2) \quad (21)$$



(a) 1^{st} channel spectrogram (time x frequency)
(b) 1^{st} predominant source transcription (time x pitch)
(c) 1^{st} non-overlapped mask (time x frequency)
(d) 2^{nd} channel spectrogram (time x frequency)
(e) 2^{nd} predominant source transcription (time x pitch)
(f) 2^{nd} non-overlapped mask (time x frequency)
(g) 3^{rd} channel spectrogram (time x frequency)
(h) 3^{rd} predominant source transcription (time x pitch)
(i) 3^{rd} non-overlapped mask (time x frequency)
(j) 4^{nd} channel spectrogram (time x frequency)
(k) 4^{nd} predominant source transcription (time x pitch)
(l) 4^{nd} non-overlapped mask (time x frequency)

Figure 4 Four channel signal overlapping mask obtention. The predominant source transcription from each channel spectrogram **(a,d,g,j)** leads to **(b,e,h,k)**. The resulting overlapping masks are displayed for each predominant source **(c,f,i,l)**.

where $|\underline{s}_j(f,t)|^2$ is called the power spectral density of source j at TF bin (f,t).

Let $\underline{x}_i(f,t)$ be the (complex-valued) STFT coefficient for mixture i at TF bin (f,t). If we assume an instantaneous mixing with coefficients $A_{i,j}$, then we can approximate the mixture spectrogram as

$$\underline{x}_i(f,t) \sim \mathcal{N}(0, \sum_j |A_{i,j}|^2 |\underline{s}_j(f,t)|^2), \tag{22}$$

where $A_{i,j}$ are the mixing matrix coefficients. In other words, the variance of $\underline{x}_i(f,t)$ is correctly modeled as $\sum_j |A_{i,j}|^2 |\underline{s}_j(f,t)|^2$. Then, each ideal separated source $s_j(t)$ can be estimated from the mixture $x_i(t)$ using a generalized time-frequency Wiener filter over the STFT domain. The Wiener filter $\alpha_{j'}$ of source j' represents the relative energy contribution of the predominant source with respect to the energy of the multichannel mixed signal $x_i(t)$ at channel i. The Wiener filter $\alpha_{j'}$ for each time-frequency bin (t,f) is defined as:

$$\alpha_{j'}(t,f) = \frac{|A_{i,j'}|^2 |\underline{s}_j(f,t)|^2}{\sum_j |A_{i,j}|^2 |\underline{s}_j(f,t)|^2}, \tag{23}$$

and the estimated source spectrogram $\hat{\underline{s}}_{j'}(t,f)$ is estimated by

$$\hat{\underline{s}}_{j'}(f,t) = \frac{\alpha_{j'}(t,f)}{|A_{i,j'}|} \underline{x}_i(t,f). \tag{24}$$

Then, the estimated source $\hat{s}_{j'}(t)$ is computed by the inverse overlap-add STFT of the estimated spectrogram $\hat{\underline{s}}_{j'}(f,t)$.

3.2.2 Separated signal decomposition

In the present work, the panning matrix is estimated and used together with the learnt instrument models to perform the separation in an NMF-based framework.

Therefore, once the panning matrix coefficients $m_{i,j}$ and the gains $g_{j,n}(t)$ are estimated, the following procedure is performed to obtain the separate sources. First of all, the estimated source magnitude spectrogram $\hat{s}_j(f,t)$ is computed with Equation 25 as follows:

$$\hat{s}_j(f,t) = g_{j,n}(t) b_{j,n}(f). \tag{25}$$

Then, the estimated Wiener mask is computed by replacing $s_j(f,t)$ by $\hat{s}_j(t,f)$ at Equation 23 as follows:

$$\hat{\alpha}_{j'}(t,f) = \frac{m_{i,j'}^2 \hat{s}_{j'}(f,t)^2}{\sum_j m_{i,j}^2 \hat{s}_j(f,t)^2}. \tag{26}$$

Then, the estimated Wiener mask for each source is applied to the multichannel signal spectrogram at channel i following Equation 24 using the phase information from the original mixture signal of the close-microphone near the target instrument. Therefore, the estimated predominant source spectrogram $\hat{s}_{j'}(f,t)$ is obtained and the estimated predominant source $\hat{s}_{j'}(t)$ is computed by applying the inverse overlap-add STFT of $\hat{\underline{s}}_{j'}(f,t)$ using the phase information from $\underline{x}_i(f,t)$ where j' is the predominant source at channel i.

4 Experiments

4.1 Training and test data

At the training stage (see Section 2.3), the basis functions are estimated using the RWC musical instrument sound database [48,49] and the full pitch range for each instrument. Four instruments are studied in the experiments (violin, clarinet, tenor saxophone, and bassoon). Individual sounds are available with a semitone frequency resolution over the entire range of notes for each instrument. Files from the RWC database have different playing styles. Files with a normal playing style and mezzo dynamic level are selected as in the literature. Training with different playing styles leads to different models. However, as demonstrated in [22], the selected configuration (normal playing style and mezzo dynamic level) is representative of the different models.

The database proposed in [27] is used for the testing stage. This database consists of ten J.S. Bach four-part chorales [27] with the corresponding aligned MIDI data. The audio files are approximately 30 s long and are sampled at 44.1 KHz from the real performances. Each music excerpt consists of an instrumental quartet (violin, clarinet, tenor saxophone, and bassoon), and each instrument is given in an isolated track.

4.2 Experimental setup

A set of synthetic recordings based on the image-source model has been generated to test the separation performance of the proposed method under a close-microphone setup. To this end, the *Roomsim* Matlab package [54] has been employed to simulate the setup depicted in Figure 5, which shows a typical source arrangement in live classical music. A large room with significantly reflective surfaces has been considered. The dimensions of the room were $22 \times 14 \times 5$ m^3 and it had a reverberation time of $T_{60} \approx 1$ s, providing a usual concert-hall acoustic environment. The sensors were configured to have cardioid directivity characteristics to reflect a usual close-miking setup. The obtained impulse responses were used to convolve the dry test signals from the evaluation database, providing the ground-truth for the different source images captured in the microphone mixture signal. These source images will be used in the performance evaluation section to compute the objective performance measures widely used by the source separation community.

4.2.1 Time-frequency representation

Many NMF-based signal processing applications usually adopt a logarithmic frequency discretization. For example, uniformly spaced subbands on the equivalent rectangular bandwidth (ERB) scale are assumed in [16,17]. In this work, two time-frequency resolutions are used. First, to estimate the instrument models and the panning matrix, a single semitone resolution was used as in [22]. In fact, the training database and the ground-truth score information are composed of notes that are separated by one semitone in frequency. This representation has proven to obtain accurate results for music

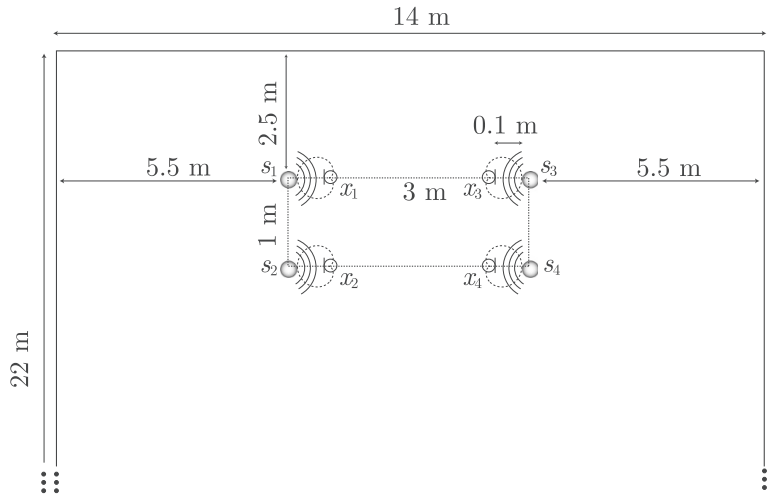


Figure 5 Simulation setup.

transcription, which is the key point when estimating the panning matrix. Second, for the separation task, a higher resolution of 1/4 of semitone is used as in [29], which has proven to achieve better separation results. These time-frequency representations are obtained by integrating the STFT bins corresponding to the same semitone, or 1/4 semitone, interval. Note that in the separation stage, the learnt basis functions $b_{j,n}(f)$ are adapted to the 1/4 semitone resolution by replicating at four times the basis at each semitone to the four samples of the 1/4 semitone resolution that belong to this semitone. The frame size and the hop size for the STFT are set to 128 and 32 ms, respectively.

4.2.2 Initialization of model parameters

The other values for the experimental parameters are the following:

1. Basis functions ($N = 88$), ranging from MIDI note 20 to 108
2. Partials per basis function for the harmonic constraint models ($M = 20$)
3. Iterations for the NMF-based algorithms (50)

The use of the beta-divergence distortion at the NMF framework makes it necessary to fit the parameter β up to the value which obtains the best results as possible. In order to find the optimum value of this parameter, a study of the separation results with different values of the parameter using the testing database has been done. At Figure 6, it is shown that the optimum value of β is around $\beta = 1.3$. This value is used for the computation of the test results.

4.2.3 Audio separation: method and metrics

For an objective evaluation of the performance of the separation method, the Perceptual Evaluation methods for Audio Source Separation (PEASS) toolkit [55] has been used. The use of objective measures based on energy ratios between the signal components, i.e., *source to distortion ratio* (SDR), the *source to interference ratio* (SIR), the *source to artifacts ratio* (SAR) and the *source image to spatial distortion ratio* (ISR), has been the standard approach in the specialized scientific community to test the quality of extracted signals.

Moreover, the overall perceptual score (OPS), the target-related perceptual score (TPS), the interference-related perceptual score (IPS) and the artifacts-related perceptual score (APS) objective measures have been used with the aim of predicting a set of subjective scores. The approach to compute the objective measures [55] makes use of auditory-motivated metrics provided by the PEMO-Q auditory model to assess the perceptual salience of the target distortion (qTarget), interference (qInterf)

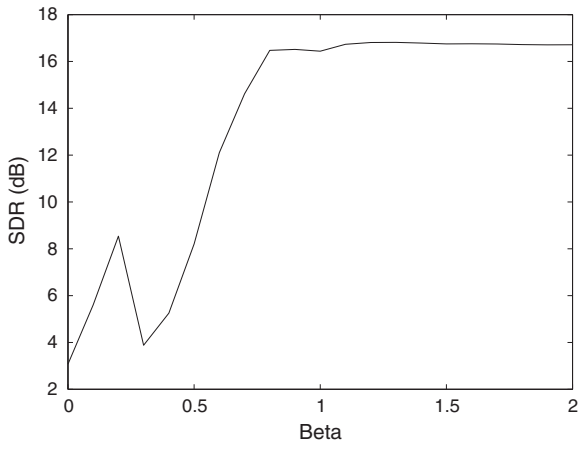


Figure 6 SDR results as a function of β.

and artifacts (qArtif), computing also a global metric (qGlobal). Then, a nonlinear mapping using neuronal networks trained with a large set of different audio signals is performed in order to get the set of objective measures. Further information about this metrics can be found in [55].

4.3 Evaluation

The proposed separation approach, shown in Figure 1, is going to be compared with some state-of-the art methods and some unrealistic situations in order to evaluate its separation capabilities. The different approaches compared here are the following:

- Default: It refers to the actual separation performed by the simulation setup presented in Figure 5. Since the sensors have cardioid directivity characteristics and they are placed following a close-miking setup, the instrument close to each microphone is predominant in the corresponding mixture channel.
- Ideal separation: This method performs as an upper bound for the best separation that can be achieved with the used time-frequency representation. The optimal value of the Wiener mask at each frequency and time component is computed assuming that the signals to be separated are known in advance.
- Oracle IMM: This approach evaluates the limitations of using an instantaneous mixing model. The separation scheme is similar to the one presented in Figure 5 but the panning matrix is estimated by knowing in advance the signals to be separated. The separation approach is identical to the proposed one but the panning matrix is optimal. We have included this method in order to evaluate the influence of the proposed panning matrix estimation stage on the final performance.
- Kokkinis method: We have included in the evaluation the results of the method proposed by Kokkinis et al. in [3] as the state-of-the-art method for addressing the microphone leakage problem.

SDR average results for all methods with the proposed testing database are presented in Figure 7. Each box represents 40 data points, one for each individual instrument of the ten mixtures test database. The lower and upper lines of each box show 25th and 75th percentiles of the sample. The line in the middle of each box is the sample mean. The lines extending above and below each box show the extent of the rest of the samples, excluding outliers that are not present in this experiment.

As can be seen, the best results are obtained with the ideal separation method, the average SDR value is about 21 dB and informs us about the best separation that can

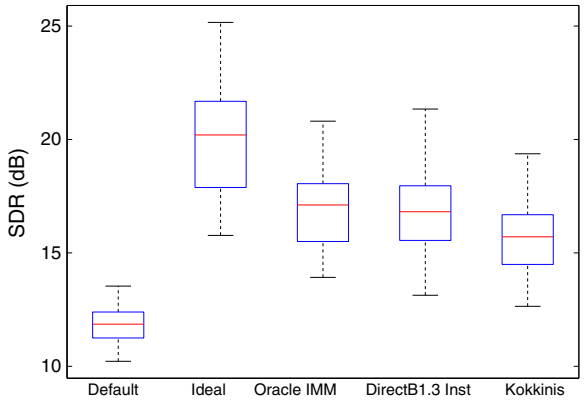


Figure 7 Comparison of the SSS methods in terms of SDR.

be achieved with the used time-frequency representation (1/4 semitone resolution in frequency). The default method is limited by the microphone leakage and obtains an average SDR of 12 dB. The oracle IMM approach provides information about the best separation results that can be obtained using our approach with an optimal panning matrix. In fact, the proposed method is similar to the Oracle solution for the studied signals; therefore, the transcription procedure in the mixture matrix stage is providing accurate estimation of the ground-truth data.

Finally, the Kokkinis' method is about 1 dB on average in SDR below the proposed method. These last results suggest that while Wiener filtering might be a simple and powerful approach to solve the microphone leakage problem, there is room for further improvements by following an informed approach such as the one presented in this paper.

To illustrate the performance of the proposed method, some separation examples can be found at https://dl.dropboxusercontent.com/u/17198775/Eurasip2013/index.html. This page contains some audio files from the test database and the corresponding sources obtained by using the methods in Table 1.

More detailed information about the separation metrics is presented in Table 1. In this table, the different metrics are given per instrument on average for the ten excerpts in the test database. In relation to the classical separation metrics (i.e., SDR, SAR, SIR and ISR, in decibels), the default method is limited by the interferences between the different instruments (SIR) while the other methods achieve better SIR results. Moreover, the bassoon separation performs worse in general, in comparison with the other instruments. This fact has been observed in other studies made by the authors [43,53]. Actually, for the bassoon case, the amplitude variations over the time line of the note cause a mismatch with the window transform in certain frequency locations (i.e., blurred regions in the spectrogram).

Table 1 Multichannel source separation results measured using PEASS toolkit

Algorithms	Inst	SDR	SAR	SIR	ISR	qTarget	qInterf	qArtif	qGlobal	OPS	TPS	IPS	APS
	Ba	12.00	23.36	12.83	24.73	0.98	0.88	0.95	0.79	36.63	78.07	56.85	29.50
Default	Cl	10.75	22.42	11.56	24.42	0.98	0.83	0.94	0.74	32.60	78.13	49.17	25.20
	Sx	11.96	23.27	12.76	25.36	0.99	0.94	0.98	0.91	41.46	66.81	60.16	63.88
	Vi	12.72	25.87	13.23	27.92	1.00	0.92	0.99	0.91	13.48	54.76	46.82	86.13
	Ba	17.03	36.49	20.13	20.49	0.98	0.96	1.00	0.96	26.78	59.50	73.25	81.87
Ideal separation	Cl	21.89	42.07	24.86	25.59	0.99	0.96	1.00	0.95	22.20	54.19	70.18	83.68
	Sx	20.11	38.82	23.18	23.75	1.00	0.98	1.00	0.98	74.20	74.40	90.10	86.74
	Vi	21.77	39.90	24.40	25.63	0.99	0.99	1.00	0.99	86.17	90.06	91.53	85.92
	Ba	14.98	35.26	15.59	25.96	0.97	0.90	1.00	0.86	9.51	39.83	28.58	69.08
Oracle IMM	Cl	16.58	34.45	16.87	29.88	0.98	0.84	1.00	0.81	8.53	38.63	16.12	77.81
	Sx	17.61	34.57	18.60	26.08	0.99	0.95	1.00	0.93	15.83	51.82	58.97	86.26
	Vi	19.28	36.23	19.81	30.79	1.00	0.95	1.00	0.95	25.76	53.67	67.97	86.36
	Ba	14.96	33.95	16.15	22.56	0.96	0.90	1.00	0.86	9.45	40.33	28.86	62.84
Directβ1.3	Cl	16.61	34.58	16.96	29.09	0.98	0.85	1.00	0.82	8.59	37.16	17.46	75.61
instant mask	Sx	16.44	33.54	19.73	20.71	0.99	0.96	1.00	0.93	15.42	56.24	57.96	82.72
	Vi	19.25	35.19	20.71	26.44	0.99	0.96	1.00	0.95	30.30	56.48	70.83	85.91
	Ba	14.02	32.53	18.65	15.77	0.99	0.88	1.00	0.87	10.41	46.35	30.90	82.24
Kokkinis [3]	Cl	16.77	34.77	19.92	19.33	0.99	0.87	0.99	0.85	9.14	49.61	24.14	83.39
	Sx	15.46	31.93	18.90	17.86	1.00	0.96	1.00	0.96	27.28	53.77	73.45	86.74
	Vi	16.58	33.40	20.77	18.52	0.99	0.96	1.00	0.95	30.75	55.54	72.55	85.92

Inst, instrument; Ba, bassoon; Cl, clarinet; Sx, tenor saxophone; Vi, violin.

Similar conclusions can be extracted by analyzing the results obtained with the perceptual similarity measures (PSM) provided by the PEMO-Q auditory model. However, with these metrics the differences between the instruments vary as a function of the analyzed signals. The Kokkinis' approach obtains better results than the proposed model (Directβ1.3) and the oracle IMM in terms of *qGlobal* (PSM values are within 0 and 1, being the unity the best result). This discrepancy between perceptual metrics (qGlobal) and classical separation metrics (SDR) can be justified after listening to the separated excerpts. In our opinion, Kokkinis' approach offers better separation capabilities at high frequencies while at low frequencies the amount of distortion seems higher. In contrast, the proposed approach seems to have higher distortion at high frequencies. Probably, the use of a softer mask in Kokkinis' approach might be a reason for these differences. It must be stressed that the optimum β value has been selected in this work in terms of SDR. Further work could be done to explore the performance of this method when the β value is selected to be optimal in terms of the perceptual metrics.

Finally, regarding the measures proposed in PEASS (i.e., OPS, TPS, IPS, and APS) there is a strong correlation between these measures and PSM. Generally, the different approaches obtain the same classification from better to worse in terms of OPS as in terms of *qGlobal* (PEASS metrics are within a 0 to 100 interval, being 100 the best result). The only exception is the default approach for the cases of bassoon and clarinet. In these cases, the OPS shows higher values, which are in contrast with the lower values of *qGlobal*. However, the separated excerpts for the default approach when playing bassoon and clarinet clearly have less perceptual quality when compared to the other approaches. The nonlinear mapping provided by the PEASS neural network does not offer satisfactory results in these cases.

5 Conclusions

In this paper, an informed NMF-based SSS method has been proposed to tackle the microphone leakage problem in multichannel close-microphone recordings. The proposed method is specifically designed for a scenario in which the number of source signals is equal or less than the number of microphone signals and a single predominant source is considered for each mixture signal. As demonstrated in the evaluation stage, despite assuming instantaneous mixing and using fixed instrument models, the proposed method provides similar performance to other state-of-the-art approaches, showing the potential

of NMF-based approaches in real-world applications. Moreover, the use of trained instrument models allows for a fast computation of the panning matrix and simplifies the separation stage by reducing the factorization to the estimation of instrument time-varying gains. However, these models are fixed and, therefore, the differences with respect to the spectra of the analyzed instruments in the mixture may lead to worse separation results, as seen in the case of the bassoon. Further work will be aimed at adapting the parameters of the model to the observed music scene. To address this issue, a proper initialization of the gains and the use of additional optimization constraints will be considered. This way, the parameters will only be adapted when there is high confidence that a note is active and free of interference.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the Andalusian Business, Science and Innovation Council under project P2010- TIC-6762, (FEDER) the Spanish Ministry of Economy and Competitiveness under the projects TEC2012-38142-C04-03 and TEC2012-37945-C02-02. The authors would like to thank the anonymous reviewers whose comments greatly helped to improve the original manuscript as well as Z. Duan for kindly sharing his annotated real-world music database.

Author details
[1]Music Technology Group (MTG), Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona 08018, Spain. [2]Computer Science Department, Universitat de València, Valencia 46010, Spain. [3]Telecommunication Engineering Department, Universidad de Jaen, Linares, Jaen 23700, Spain.

Received: 29 June 2013 Accepted: 21 November 2013
Published: 13 December 2013

References

1. DM Huber, RE Runstein, *Modern Recording Techniques*, 7th edn. (Focal Press, UK, 2009)
2. A Clifford, JD Reiss, Microphone interference reduction in live sound, in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)* (Paris, 19–23 September 2011)
3. EK Kokkinis, J Mourjopoulos, Unmixing acoustic sources in real reverberant environments for close-microphone applications. J. Audio Eng. Soc. **58**(11), 907–922 (2010)
4. EK Kokkinis, JD Reiss, J Mourjopoulos, A Wiener filter approach to microphone leakage reduction in close-microphone applications. IEEE Trans. Audio, Speech, Language Process. **20**(3), 767–779 (2012)
5. P Comon, C Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. (Academic Press, Oxford, 2010)
6. DD Lee, HS Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**, 799–791 (1999)
7. O Yilmaz, S Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
8. M Cobos, JJ Lopez, Maximum a posteriori binary mask estimation for underdetermined source separation using smoothed posteriors. IEEE Trans. Audio, Speech, Language Process. **20**(7), 2059–2012 (2012)
9. MS Pedersen, J Larsen, U Kjems, LC Parra, Convolutive blind source separation methods, in *Springer Handbook of Speech Processing*, ed. by J Benesty, MM Sondhi, and Y Huang (Springer, Berlin, 2008), pp. 1065–1084
10. P Smaragdis, Blind separation of convolved mixtures in the frequency domain. Neurocomputing **22**(1), 21–34 (1998)
11. T Virtanen, Sound source separation in monaural music signals, Thesis, Tampere University of Technology, 2006
12. P Smaragdis, J Brown, Non-negative matrix factorization for polyphonic music transcription, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, 19–22 October 2003)
13. M Cobos, P Vera-Candeas, JJ Carabias-Orti, N Ruiz-Reyes, JJ Lopez, Blind estimation of reverberation time from monophonic instrument recordings based on non-negative matrix factorization, in *Proceedings of the AES 42nd International Conference: Semantic Audio* (Ilmenau, 22–24 July 2011)
14. A Ozerov, C Févotte, Multichannel non-negative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. Audio, Speech, Language Process. **18**(3), 550–563 (2010)
15. R Hennequin, R Badeau, B David, Time-dependent parametric and harmonic templates in non-negative matrix factorization, in *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (Graz, 6–10 September 2010), pp. 246–253
16. E Vincent, N Bertin, R Badeau, Adaptive harmonic spectral decomposition for multiple pitch estimation. IEEE Trans. Audio, Speech, Language Process. **18**(3), 528–537 (2010)
17. N Bertin, R Badeau, E Vincent, Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. IEEE Trans. Audio, Speech, Language Process. **18**(3), 538–549 (2010)
18. K Itoyama, M Goto, K Komatani, T Ogata, HG Okuno, Instrument equalizer for query-by-example retrieval: improving sound source separation based on integrated harmonic and inharmonic models, in *Proceedings of the International Conference for Music Information Retrieval (ISMIR)* (Philadelphia, 14–18 September 2008), pp. 133–138
19. J Wu, E Vincent, SA Raczynski, T Nishimoto, N Ono, S Sagayama, Multipitch estimation by joint modeling of harmonic and transient sounds, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Prague, 22–27 May 2011), pp. 25–28
20. T Heittola, A Klapuri, T Virtanen, Musical instrument recognition in polyphonic audio using source-filter model for sound separation, in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (Kobe, 26–30 October 2009), pp. 327–332
21. JL Durrieu, B David, G Richard, A musically motivated mid-level representation for pitch estimation and musical audio source separation. IEEE J. Selected Topics Signal Process. **5**(6), 1180–1191 (2011)
22. JJ Carabias-Orti, T Virtanen, P Vera-Candeas, N Ruiz-Reyes, FJ Cañadas-Quesada, Musical instrument sound multi-excitation model for non-negative spectrogram factorization. IEEE J. Selected Topics Signal Process. **5**(6), 1144–1158 (2011)
23. A Ozerov, A Liutkus, R Badeau, G Richard, Informed source separation: source coding meets source separation, in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)* (New Paltz, 16–19 October 2011)
24. JJ Bosch, K Kondo, R Marxer, J Janer, Score-informed and timbre independent lead instrument separation in real-world scenarios, in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (Bucharest, 27–31 August 2012), pp. 2417–2421
25. R Hennequin, B David, R Badeau, Score informed audio source separation using a parametric model of non-negative spectrogram, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Prague, 22–27 May 2011), pp. 45–48
26. J Ganseman, P Scheunders, G Mysore, J Abel, Evaluation of a score-informed source separation system, in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (Utrecht, 9–13 August 2010)
27. Z Duan, B Pardo, Soundprism: an online system for score-informed source separation of music audio. Selected Topics Signal Process. IEEE J. **5**(6), 1205–1215 (2011)
28. U Simsekli, AT Cemgil, Score guided musical source separation using generalized coupled tensor factorization, in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, (27–31 August 2012), pp. 2639–2643
29. FJ Rodriguez-Serrano, JJ Carabias-Orti, P Vera-Candeas, FJ Canadas-Quesada, N Ruiz-Reyes, Monophonic constrained non-negative sparse coding using instrument models for audio separation and transcription of monophonic source-based polyphonic mixtures.

Multimedia Tools Appl (2013). Available at http://link.springer.com/article/10.1007%2Fs11042-013-1398-8
30. B Fuentes, R Badeau, G Richard, Blind harmonic adaptive decomposition applied to supervised source separation, in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (Bucharest, 27–31 August 2012), pp. 2654–2658
31. A Hurmalainen, J Gemmeke, T Virtanen, Detection, separation and recognition of speech from continuous signals using spectral factorisation, in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (Bucharest, 27–31 August 2012), pp. 2649–2653
32. D Fitzgerald, User assisted separation using tensor factorisations, in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (Bucharest, 27–31 August 2012), pp. 2412–2416
33. A Liutkus, J Pinel, R Badeau, L Girin, G Richard, Informed source separation through spectrogram coding and data embedding. Signal Process. **92**(8), 1937–1949 (2012)
34. M Casey, A Westner, Separation of mixed audio sources by independent subspace analysis, in *Proceedings of the International Computer Music Conference (ICMC '00)* (Berlin, September 2000), pp. 154–161
35. T Virtanen, Sound source separation using sparse coding with temporal continuity objective, in *Proceedings of the International Computer Music Conference (ICMC '03)* (Singapore, September 2003)
36. T Virtanen, Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech, Language Process. **15**(3), 1066–1074 (2007)
37. L Benaroya, F Bimbot, R Gribonval, Audio source separation with a single sensor. Audio, Speech, Language Process. IEEE Trans. **14**(1), 191–199 (2006)
38. AT Cemgil, P Peeling, O Dikmen, S Godsill, Prior structures for time-frequency energy distributions, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, 21–24 October 2007), pp. 151–154
39. A Liutkus, R Badeau, G Richard, Gaussian processes for underdetermined source separation. Signal Process. IEEE Trans. **59**(7), 3155–3167 (2011)
40. T Virtanen, A Klapuri, Analysis of polyphonic audio using source-filter model and non-negative matrix factorization, in *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop* (Whistler, 9 December 2006)
41. SA Raczyński, N Ono, S Sagayama, Multipitch analysis with harmonic nonnegative matrix approximation, in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)* (Vienna, 23–27 September 2007), pp. 381–386
42. D FitzGerald, M Cranitch, E Coyle, Extended nonnegative tensor factorisation models for musical source separation. Comput. Intell. Neurosci. **2008**, 15 (2008). Article ID 872425
43. JJ Carabias-Orti, FJ Rodriguez-Serrano, P Vera-Candeas, FJ Canadas-Quesada, N Ruiz-Reyes, Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription. Eng. Appl. Artif. Intell. **26**(7), 1671–1680 (2013)
44. A Ozerov, E Vincent, F Bimbot, A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. Audio, Speech, Language Process. **20**(4), 1118–1133 (2012)
45. C Févotte, N Bertin, JL Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
46. C Févotte, J Idier, Algorithms for nonnegative matrix factorization with the beta-divergence. Neural Comput. **23**(9), 242–2456 (2011)
47. DD Lee, HS Seung, Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems. **13**, 556–562 (2001)
48. M Goto, H Hashiguchi, T Nishimura, R Oka, RWC music database: popular, classical, and jazz music databases, in *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR)* (Paris, 13–17 October 2002)
49. M Goto, Development of the RWC music database, in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)* (Kyoto, 4–9 April 2004), pp. I-553-556. (invited paper)
50. RM Parry, IA Essa, Estimating the spatial position of spectral components in audio, in *Proceedings of the 6th International Conference of Independent Component Analysis and Blind Signal Separation (ICA'06)* (Charleston, 5–8 March 2006), pp. 666–673
51. D FitzGerald, M Cranitch, E Coyle, Non-negative tensor factorisation for sound source separation, in *Proceedings of the Irish Signals and Systems Conference* (Dublin, September 2005), pp. 8–12
52. C Fevotte, A Ozerov, Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues, in *Proceedings of the 7th International Symposium on Computer Music Modeling and Retrieval (CMMR)* (Malaga, 21–24 June 2010), pp. 102–115
53. JJ Carabias-Orti, T Virtanen, P Vera-Candeas, N Ruiz-Reyes, FJ Canadas-Quesada, Musical instrument sound multi-excitation model for non-negative spectrogram factorization. IEEE J. Selected Topics Signal Process. **5**(6), 1144–1158 (2011)
54. DR Campbell, KJ Palomaki, GJ Brown, A MATLAB simulation of "shoebox" room acoustics for use in research and teaching. Comput. Inf. Syst. J. **9**(3), 48–51 (2005)
55. V Emiya, E Vincent, N Harlander, V Hohmann, Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio, Speech Language Process. **19**(7), 2046–2057 (2011)

doi:10.1186/1687-6180-2013-184
Cite this article as: Carabias-Orti *et al.*: **Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings.** *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:184.