This article was downloaded by: [Biblioteca, Universitat Pompeu Fabra]
On: 20 May 2015, At: 04:23
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

CrossMark

Click for updates

Journal of New Music Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nnmr20

Online Score-Informed Source Separation with Adaptive Instrument Models

Francisco J. Rodriguez-Serrano[a], Zhiyao Duan[b], Pedro Vera-Candeas[a], Bryan Pardo[c] & Julio J. Carabias-Orti[d]

[a] Department of Telecommunication Engineering, University of Jaen, Spain.

[b] Department of Electrical and Computer Engineering, University of Rochester, NY, USA.

[c] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA.

[d] Music Technology Group Universitat Pompeu Fabra, Barcelona.

Published online: 27 Jan 2015.

To cite this article: Francisco J. Rodriguez-Serrano, Zhiyao Duan, Pedro Vera-Candeas, Bryan Pardo & Julio J. Carabias-Orti (2015): Online Score-Informed Source Separation with Adaptive Instrument Models, Journal of New Music Research, DOI: 10.1080/09298215.2014.989174

To link to this article: http://dx.doi.org/10.1080/09298215.2014.989174

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Journal of New Music Research, 2015
http://dx.doi.org/10.1080/09298215.2014.989174

Routledge
Taylor & Francis Group

Online Score-Informed Source Separation with Adaptive Instrument Models

Francisco J. Rodriguez-Serrano[1], Zhiyao Duan[2], Pedro Vera-Candeas[1], Bryan Pardo[3] and Julio J. Carabias-Orti[4]

[1]*Department of Telecommunication Engineering,University of Jaen, Spain;* [2]*Department of Electrical and Computer Engineering, University of Rochester, NY, USA;* [3]*Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA;* [4]*Music Technology Group Universitat Pompeu Fabra, Barcelona*

(Received 12 March 2014; accepted 10 November 2014)

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Abstract

In this paper, an online score-informed source separation system is proposed under the Non-negative Matrix Factorization (NMF) framework, using parametric instrument models. Each instrument is modelled using a multi-excitation source-filter model, which provides the flexibility to model different instruments. The instrument models are initially learned on training excerpts of the same kinds of instruments, and are then adapted, during the separation, to the specific instruments used in the audio being separated. The model adaptation method needs to access the musical score content for each instrument, which is provided by an online audio-score alignment method. Source separation is improved by adapting the instrument models using score alignment. Experiments are performed to evaluate the proposed system and its individual components. Results show that it outperforms a state-of-the-art comparison method.

Keywords: NMF, online, score-informed, instrument-models, adaptive, score alignment, source separation

1. Introduction

The goal of Sound Source Separation (SSS) is to segregate constituent sound sources from an audio signal mixture. SSS enables all kinds of users, from amateurs up to professionals, to work with separated sources. The SSS task is of interest because a lot of direct user applications can be developed with it. Personalizing a live concert by letting the listener adjust the volume of individual instruments is one application of online SSS. Music education applications can be developed with both offline and online SSS methods. For example, the sound of one instrument can be removed from the recording, so that a live musician could perform the removed part, using the recording as an accompaniment. Also, a rehearsal of several musicians could be recorded and separated online, at the end of the rehearsal they would have the separated recordings in order to check for mistakes and improve their performance. The separated sources would be available without the need to record them with a multichannel system and a specialized studio. Also, SSS is a useful preprocessing stage for other research tasks, such as automatic music transcription (Gainza & Coyle, 2007), structured coding (Viste & Evangelista, 2001) and beat tracking (Chordia & Rae, 2009). Using separated sources simplifies all these tasks, improving their results even when source separation is not perfect.

Depending on the number of sources (musical instruments) and sensors (microphones) used in the mixed signal recording, the SSS problem can be classified into three cases. *Overdetermined* cases are those where the number of sensors is larger than the number of sources (Hyvarinen & Oja, 2000; Zibulevsky, Kisilev, & Zeevi, 2002). In *determined* cases the number of sensors and sources is the same. Finally, in *underdetermined* cases, there are more sources than sensors. The *overdetermined* and *determined* cases are usually addressed with methods (e.g. Independent Component Analysis or Independent Subspace Analysis) that cannot be applied when the number of sources exceeds the number of sensors, since they depend on having at least as many sensors as sound sources (Babaie-zadeh & Jutten, 2006). The underdetermined case is the most common one for music recordings and performances (e.g. mono or stereo mixtures of three or more instruments or voices). An important and commonly used framework for addressing the underdetermined case is Non Negative Matrix Factorization (NMF) (Bryan et al., 2000; Virtanen & Klapuri, 2006).

Depending on the use, or not, of prior information, the SSS task is called *Informed Source Separation* (ISS), or *Blind*

Correspondence: F.J. Rodriguez-Serrano, EPSL Alfonso X El Sabio 28, 23700 Linares, Jaen, Spain. E-mail: fjrodrig@ujaen.es

© 2015 Taylor & Francis

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Source Separation (BSS) (Comon & Jutten, 2010). To date, the performance of BSS is very dependent on the signal nature and does not reliably achieve good enough quality on the separated sources for practical use in music applications. Instead, the way to obtain robust separation in practice is to use ISS. There are several types of information that can be used in ISS. Spectral information can be introduced by using instrument models when the instruments are known in advance (Ewert & Muller, 2012; Fritsch & Plumbley, 2013; Rodriguez-Serrano, Carabias-Orti, Vera-Candeas, Canadas-Quesada, & Ruiz-Reyes, 2013; Simsekli & Cemgil, 2012). Also, musical score information can be used if the score and audio are well aligned (Duan & Pardo, 2011; Ewert & Muller, 2012; Fritsch & Plumbley, 2013; Ganseman, Scheunders, Mysore, & Abel, 2010; Hennequin, David, & Badeau, 2011). In this paper we deal with the problem of online score-informed separation of harmonic musical sources from a single-channel recording. We combine the audio-score alignment model proposed in Duan and Pardo (2011) with the Multi-excitation per Instrument (MEI) NMF model proposed in Carabias-Orti, Virtanen, Vera-Candeas, Ruiz-Reyes, and Cañadas-Quesada (2011) to build our baseline system. We then advance this system by (1) adapting the pre-learned instrument models towards the real instruments played in the music with the information of score alignment, and (2) designing online algorithms for the instrument adaptation and source separation. Therefore, our final system takes a music score and pre-learned instrument models as prior information, aligns the score with the audio, updates the instrument models towards the real played instruments, and separates the audio mixture, all completed in an online fashion.

In the experiments, we show that the use of instrument models and the adaptation of these models to the instruments used in the musical performance, significantly improves the source separation performance. We also show that the proposed online algorithm separates sources almost as well as the offline version of the algorithm.

1.1 Related work

Ewert and Muller (2012) proposed a system that initializes the instrument models (spectral patterns) of different instruments as a harmonic comb with a constant declining amplitude, and adapts the models to real played instruments. While this does not require a pre-learning phase of instrument models as in our proposed system, the initialization may not fit well to instruments that have many fluctuations in the harmonics such as clarinet and bassoon.

Fritsch and Plumbley (2013) presented a method for musical audio source separation, using the information from the musical score to supervise the decomposition process based on an NMF framework. They initialize the instrument models with those pre-learned from a MIDI-synthesized audio, and then adapt them to the real played instruments in the mixture. This initialization is closer to the real played instruments than the harmonic comb initialization in Ewert and Muller (2012), but it depends heavily on the synthesizer.

The main features of our proposed method that distinguish it from the adaptation and separation approaches in Ewert and Muller (2012), and Fritsch and Plumbley (2013) are: (1) the initial instrument models are learned from real instrument recordings, instead of artificial templates or synthesized audio; (2) the adaptation of our models is performed using only non-overlapped partials, identified using score information (this makes the method more robust in polyphonic scenarios); and (3) our method only requires access to current and past audio frames when separating the current audio frame (i.e. works online), while both Fritsch and Plumbley (2013), and Ewert and Muller (2012) require access to future audio frames (i.e. are offline methods).

In this paper, we use the term algorithmic latency as the delay between receiving the signal and starting to perform separation. In our approach, the algorithmic latency is half a frame, because we start to align and separate a frame right after we receive it. In practice, the real latency depends not only on the algorithmic latency but also on the implementation. However, the latency introduced by the implementation can be improved by using more advanced computers or more optimized programming while the algorithmic latency cannot.

There are online approaches for source separation under the NMF-SSS framework (Duan, Mysore, & Smaragdis, 2012; Joder, Weninger, Eyben, Virette, & Schuller, 2012; Simon & Vincent, 2012). However, Duan et al. (2012), and Joder et al. (2012) were only tested in speech enhancement applications as they were designed to adapt one source (speech or noise) but keep the other source fixed during separation. In our proposed method, instrument models of multiple instruments of the music are adapted simultaneously. The method proposed in Simon and Vincent (2012) is suitable for multichannel signals, but not for monaural ones. The mixing information (which represents the spatial information) is very important for their system. Also, random initialization of the model parameters without any extra information would make it difficult to discriminate between the different sources at monaural sources. To date, none of these approaches (Duan et al., 2012; Joder et al., 2012; Simon & Vincent, 2012) have been applied to work in the score-informed source separation setting.

Besides Duan and Pardo (2011), there are several other online polyphonic audio-score alignment methods (Cont, 2006, 2006; Dixon and Widmer, 2005). We use the method proposed in Duan and Pardo (2011) because it is designed to align multi-instrument polyphonic music audio with score information and it is has been tested on multi-instrumental polyphonic audio with clear objective measures. Despite the fact that the proposal of Dixon and Widmer (2005) is tested over piano music and multi-instrument signals, its performance is not evaluated over multi-instrument signals with objective results, due to the lack of reliable annotations. Other methods are only designed for (Cont, 2006), or tested on (Cont, 2010), single-instrument polyphonic audio.

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

The rest of the paper is structured as follows. Section 2 reviews the background that the proposed methods are built on. Section 3 describes the proposed methods. Experiments and the comparison to other state-of-the-art methods are described in Section 4. Finally, we draw conclusions and discuss future work in Section 5.

2. Background

In this section the methods from the bibliography that the proposed system builds on are summarized. The aim of the proposed work is to implement an *Online Score-Informed Source Separation System with Adaptative Instrument Models*. It is uses an NMF framework initialized with the score information for the activations and previously trained instrument models for the spectral patterns. This framework, suitably modified, should be able to update these models while factorizing the signal in an online manner.

2.1 Audio-score alignment

The audio-score alignment module of the proposed score-informed source separation method was proposed in Duan and Pardo (2011). It is an online algorithm that aligns a score to a piece of polyphonic music audio played by multiple instruments. The basic idea is to view an audio performance as a path in a two-dimensional state space, where the two dimensions are score position and tempo, respectively. The state space is continuous and the path is hidden. The aim is to infer this path from the observed audio signal in an online fashion.

Mathematically, the n-th time frame of the audio performance is represented as $\mathbf{y}_n$, and is associated with a two-dimensional state variable $\mathbf{s}_n = (x_n, v_n)^T$, where x_n is its score position (in beats), v_n is its tempo (in beats-per-minute (BPM)) and T denotes the matrix transposition. The aim is to infer the current score position x_n from current and previous observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$. This problem is modelled by a hidden Markov process.

A hidden Markov process model contains two parts: a process model $p(\mathbf{s}_n|\mathbf{s}_{n-1})$ that describes how the states transition from one to another, and an observation model $p(\mathbf{y}_n|\mathbf{s}_n)$ that describes the likelihood of a state s_n generating the observation $\mathbf{y}_n$. The difference of a hidden Markov process from a finite-state hidden Markov model is that the states are continuous variables and they can take infinitely many values.

The process model $p(\mathbf{s}_n|\mathbf{s}_{n-1})$ is defined through two dynamic equations. The score position changes from the previous position according to the tempo. The tempo changes through a random walk or does not change at all, depending on where the position is. The observation model $p(\mathbf{y}_n|\mathbf{s}_n)$ is defined through the multi-pitch estimation likelihood model proposed in Duan, Pardo and Zhang (2010). For any set of pitches, the multi-pitch estimation likelihood model computes its likelihood to fit the observed audio frame. Now, given a hypothesized state, the set of pitches that are supposed to be played in the n-th audio frame can be read from the score at the score position of the state. By plugging the set of score pitches into the multi-pitch estimation likelihood model, one can calculate the likelihood this set of pitches would result in the observed audio frame. The higher the likelihood is, the better fit the set of pitches has, and the better the state hypothesis is. Given the process model and the observation model, Duan and Pardo (2011) use particle filtering to infer the hidden states from the observations, one frame after another, in an online fashion.

2.2 Signal factorization with MEI

2.2.1 Multi-excitation per instrument (MEI)

Let $X(t, f)$ be the true time–frequency representation of an audio mixture (e.g. a recording of several musical instruments), where t is time and f is a frequency of analysis. Let $\hat{X}(t, f)$ be an estimate of the true mixture. We define a spectral basis function $b(f)$ as a function that outputs the relative amplitudes of each frequency. We use spectral basis functions to represent the instantaneous timbre of sound sources, like musical instruments. If the timbre of an instrument is different in different situations (e.g. when a musical instrument plays different pitches) multiple spectral basis functions (one per pitch) can be associated with the instrument.

The MEI framework tries to decompose $\hat{X}(t, f)$ of the audio mixture into a linear combination of spectral basis function:

$$X(t, f) \approx \hat{X}(t, f) = \sum_{j=1}^{J} \sum_{n=1}^{N} g_{n,j}(t) b_{n,j}(f), \qquad (1)$$

where $b_{n,j}(f)$ is the n-th basis for the j-th instrument; $g_{n,j}(t)$ is its gain at frame t. When dealing with harmonic instrument sounds in this paper, each spectral basis function ideally corresponds to a pitch, and the gain represents the activation strength of the pitch.

The multi-excitation model proposed by Carabias-Orti et al. (2011) is an extension of the regular excitation-filter model presented in Virtanen and Klapuri (2006). The regular excitation-filter model has origins in speech processing and sound synthesis. In speech processing, the excitation models the sound produced by the vocal cords, whereas the filter models the resonating effect of the vocal tract (Rabiner & Schafer, 1978). In sound synthesis, excitation-filter (or source-filter) synthesis (Välimäki, Pakarinen, Erkut, & Karjalainen, 2006) colours a spectrally rich excitation signal to get the desired sound.

The spectral basis functions $b_{n,j}(f)$ in Equation 1 depend on both pitch and instrument, and this results in a large number of functions, where the value at each frequency for each function must be estimated. To reduce the number of values that must be learned, Virtanen and Klapuri (2006) model each basis function $b_{n,j}(f)$ as the product of a pitch-dependent excitation spectrum $e_n(f)$ and an instrument-dependent filter $h_j(f)$:

$$b_{n,j}(f) = h_j(f) e_n(f), \quad n = 1, \ldots, N, \; j = 1, \ldots, J. \qquad (2)$$

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

$e_n(f)$ encodes the pitch information. $h_j(f)$ encodes the frequency response of the resonance body of the instrument. This approach significantly reduces the number of parameters. However, since a piece of music can contain many different pitches and for each pitch a full spectrum is needed to represent $e_n(f)$, there are still many parameters to tune.

To further reduce the need to learn values from data, several studies (Badeau, Emiya, & Emiya, 2009; Klapuri, Virtanen, & Heittola, 2010; Heittola, Klapuri, & Virtanen, 2009) introduce the excitations $e_n(f)$ as frequency components of unity magnitude at integer multiples of the fundamental frequency of the pitch n. This results in modelling the spectral basis functions as the product of an instrument-dependent filter and a *harmonic comb* excitation, where each component of the comb is a shifted frequency response of the window function:

$$b_{n,j}(f) = \sum_{m=1}^{M} h_j(mf_0(n))G(f - mf_0(n)), \qquad (3)$$

where M is the number of harmonics, $f_0(n)$ is the fundamental frequency of pitch n, and $G(f)$ is the magnitude spectrum of the window function. The $G(f - mf_0(n))$ term is included because a harmonic constraint is imposed. This harmonic constraint considers that harmonic signals are produced as a sum of harmonic-related tones.

The above unity-magnitude harmonic comb excitation model, together with the instrument-dependent filter (which often has a smooth frequency response), is able to represent some instruments with a smooth envelope of their spectral peaks. However, the spectral envelope of other instruments, such as the clarinet, are not smooth and they cannot be well represented with a flat excitation function. For example, the second harmonic of a clarinet note is often very soft, no matter what pitch the note has. This makes it impossible to represent the spectral envelopes of different clarinet notes with a single filter. More details can be obtained in Carabias-Orti et al. (2011) where examples for different instruments and pitches are presented.

An interesting alternative is the use of the multi-excitation model proposed in Carabias-Orti et al. (2011). This model defines the excitation spectrum as a linear combination of a few excitation basis vectors. In fact, the regular excitation-filter model presented in Equation 2 requires an excitation per pitch (per instrument) $e_n(f)$, while the multi-excitation model just requires as few as $I = 2$ excitations per instrument to properly model the instruments from their test database (Carabias-Orti et al., 2011).

Under the multi-excitation model, the excitation per pitch and instrument is defined as

$$e_{n,j}(f) = \sum_{m} \left(\sum_{i} w_{i,n,j} v_{i,m,j} \right) G(f - mf_0(n)), \qquad (4)$$

where $m = 1, ..., M$ is the index of the harmonics; and $i = 1, ..., I$ is the index of excitation basis vectors ($I \ll N$). $v_{i,m,j}$ is the m-th harmonic of the i-th excitation basis vector for instrument j; $w_{i,n,j}$ is the weight of the i-th excitation basis vector for pitch n and instrument j. $G(f - mf_0(n))$ is the window transform placed at the frequency of the m-th harmonic of the n-th pitch.

The key of the MEI model is the separation of the excitation $e_{n,j}(f)$ into two parts: the excitation basis vectors $v_{i,m,j}$ and their weights $w_{i,n,j}$. For each instrument j, there are I excitation basis vectors of dimension M. The amplitude of each partial m of the final excitations $e_{n,j}(f)$ is a linear combination of I excitation basis vectors $v_{i,m,j}$ weighted by $w_{i,n,j}$. The excitation basis vectors are instrument dependent but are not pitch dependent. The weights in the linear combination, however, are both instrument dependent and pitch dependent. The excitation basis vector contains 20 partials and two excitation basis vectors are used for each instrument. There are N pairs of weights, one pair for each pitch index. The weights are the scalar values that multiplies each excitation basis vector and they are linearly combined to obtain the 20 excitation partials for the concrete pitch. The spectral patterns are finally obtained with the multi-excitation model as

$$b_{n,j}(f) = \sum_{i,m} h_j(mf_0(n)) w_{i,n,j} v_{i,m,j} G(f - mf_0(n)). \qquad (5)$$

All these parameters of the MEI model are summarized at Table 1.

Different instruments have different pitch ranges. For example, the pitch range of the bassoon covers 37 semitones while the pitch range of the violin covers 45 semitones. In this paper, we use the resolution of 1/8 semitones for the pitch indexes. Therefore, the number of the pitch indexes N for bassoon is $37 \times 8 = 296$ semitones while that for violin is $45 \times 8 = 360$ semitones.

While audio is typically encoded in a time–frequency representation using linearly-spaced frequency bins, we convert the input Short Time Fourier Transform (STFT) with linear-frequency into a log-frequency representation. This lets us use the same 1/8 semitone-spaced indexes for both the frequency indexes F and the pitch indexes N used to build our spectral pattern models. There are two reasons that we use the same resolution for frequency indices and pitch indices. First, the spectral pattern (frequency) of a basis function (pitch) is not sensitive to small deviations of the played pitch. In other words, two different sounds of the same instrument and pitch can be very different in linear frequency with a small deviation in logarithmic pitch (specially at high frequencies) but not so different when a logarithmic frequency resolution is used. Second, the dimensionality of the spectrogram is reduced and the computational complexity of factorization is significantly decreased. The use of the log-frequency representation and the selected analysis parameters (sample rate of 44,100 Hz, window size of 128 ms and STFT of 8192 points) results in some low frequency indexes without associated frequency bins in the STFT. These indexes are not used. This results in a frequency index with 608 values, which covers from the lowest frequency generated by the tested instrument up to 22,050 Hz.

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Table 1. Parameters and their sizes of the MEI signal model.

Parameter	Size	Description
$X(t, f)$	$F \times T$	Time/Frequency signal input
$\hat{X}(t, f)$	$F \times T$	Time/Frequency signal reconstruction
$g_{n,j}(t)$	$N \times T \times J$	Gain for each pitch and instrument at each frame
$h_j(f)$	$F \times J$	Filter for each instrument at each frequency
$v_{i,m,j}$	$I \times M \times J$	Excitation basis vectors. I vectors with M values for each instrument
$w_{i,n,j}$	$I \times N \times J$	Weight of each excitation basis vector for each pitch and instrument
$b_{n,j}(f)$	$N \times F \times J$	Basis functions based on $h_j(f)$, $w_{i,n,j}$ and $v_{i,m,j}$, as described in Equation 5
$G(f - mf_0(n))$	$F \times 1$	Spectrum of the analysis window placed at the $mf_0(n)$ frequency
F	608	Number of frequency indexes with 1/8 semitones as maximum resolution
T	–	Number of analysis frames
N	–	Instrument-dependent number of pitch indexes (1/8 semitones resolution)
J	–	Number of sources (instruments)
M	20	Number of harmonics considered per excitation
I	2	Number of excitation basis vectors per instrument

The MEI model parameters reduction can be demonstrated with a simple example. The total number of parameters with the MEI model, the regular excitation-filter model and the harmonic comb excitation version are calculated for the clarinet case. The clarinet can play 37 semitones on the chromatic scale, which correspond to 296 pitch indexes ($N = 296$) with the 1/8 semitone resolution used in this work. The source filter model needs $608 \times 1 = 608$ parameters ($F \times J$) for the filter ($h_j(f)$) and $296 \times 20 = 5920$ parameters ($N \times M$) for the excitation basis vectors ($e_n(f)$), which becomes a total 6528 parameters for the clarinet source filter model. In the case of the harmonic comb excitation, the excitation parameters are not required and only 608 parameters are needed for the filter component. According to Table 1, the MEI model needs the same 608 parameters for representing the filter ($h_j(f)$), $2 \times 20 \times 1 = 40$ parameters ($I \times M \times J$) for representing the excitation basis vectors ($v_{i,m,j}$) and $2 \times 296 \times 1 = 592$ parameters ($I \times N \times J$) for representing the weights of the excitation basis vectors ($w_{i,n,j}$). This means that MEI needs 1240 parameters to represent a clarinet model.

The lightest model is the harmonic comb excitation, however its flat excitation component is not able to represent a non-smooth spectral envelope, as in the case of the clarinet (Carabias-Orti et al., 2011). On the other hand, the regular excitation-filter and MEI models both correctly represent the spectral behaviour of different instruments. In the given example, the MEI model reduces the number of parameters used to represent a clarinet model by 77%, compared to the regular excitation-filter model. To summarize, the MEI model preserves the flexibility of the regular excitation-filter model with a lower number of parameters.

Given the MEI model, the magnitude spectra of the mixture signal can be decomposed by substituting Equation 5 into Equation 1:

$$\hat{X}(t, f) = \sum_{n,m,i,j} g_{n,j}(t) h_j(mf_0(n)) w_{i,n,j} v_{i,m,j} G(f - mf_0(n)). \tag{6}$$

Given the NMF model in Equation 6, we want to estimate the parameters so that the reconstruction error between the observed spectrogram $X(t, f)$ and the modelled one $\hat{X}(t, f)$ is minimized. The β-divergence (Fevotte & Idier, 2011; Vincent, Bertin, & Badeau, 2010) is used here as the cost function to define the reconstruction error, where β is in the range of [0, 2]. The use of the β-divergence as distortion measure makes the system flexible and it allows one to study its behaviour with the most used divergences (Euclidean or $\beta = 2$, Kullback–Leibler or $\beta = 1$ and Itakura–Saito or $\beta = 0$). In Carabias-Orti et al. (2011) the MEI model is employed only with the Kullback–Leibler divergence, which corresponds to the case $\beta = 1$. In Fritsch and Plumbley (2013) and Hennequin et al. (2011) β-divergence is used, but they set $\beta = 1$ which is the same as using KL divergence. In Section 4, a study of the separation performance in function of the parameter β is shown.

In Lee and Seung (2001), an iterative algorithm based on multiplicative update rules is proposed to obtain the model parameters that minimize the cost function. Under these rules, $D_\beta(X_t(f) \| \hat{X}_t(f))$ is non-increasing at each iteration and the non-negativity of the bases and the gains is ensured. The multiplicative update rule (see Lee and Seung (2001) for further details) for each scalar parameter θ_l is given by expressing the partial derivatives of the $\nabla_{\theta_l} D_\beta$ as the quotient of two positive terms $\nabla^-_{\theta_l} D_\beta$ and $\nabla^+_{\theta_l} D_\beta$:

$$\theta_l \leftarrow \theta_l \frac{\nabla^-_{\theta_l} D_\beta(X(t, f) || \hat{X}(t, f))}{\nabla^+_{\theta_l} D_\beta(X(t, f) || \hat{X}(t, f))}, \tag{7}$$

assuming $\nabla_{\theta_l} D = \nabla^+_{\theta_l} D - \nabla^-_{\theta_l} D$. The main advantage of the multiplicative update rule in Equation 7 is that non-negativity of the bases and the gains is ensured, resulting in a non-negative matrix factorization (NMF) algorithm.

There are other proposals in the literature to address the factorization problem. For instance, Alternative Least Squares (ALS) is proposed by Cichocki and Zdunek (2007). Also, Gillis and Luce (2014) propose an optimization of the

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

factorization process by linear programming. Here we adopt the mostly widely used Lee and Seung algorithm.

3. The proposed method

In this section, we describe the proposed online adaptive score-informed source separation method. To make the presentation clear, we present the method in four steps, each built upon the previous step. In Section 3.1 the training process is described. In Section 3.2, we present a basic source separation algorithm, where the instrument models are learned from the training data and kept fixed during separation. In Section 3.3, we propose to adapt the trained instrument model to the real instrument in the audio mixture. In Section 3.4, we make the separation and adaptation online.

3.1 Instrument modelling

The model reviewed in Section 2.2.1 requires an estimate of the spectral basis functions $b_{n,j}(f)$ for each note n and instrument j. Each function $b_{n,j}(f)$ is learned in advance using isolated notes of solo instrument recordings in the RWC database (Goto, 2004; Goto, Hashiguchi, Nishimura, & Oka, 2002) (for more details see the experimental setup section). We initialize each gain $g_{n,j}(t)$ from the ground-truth pitch transcription of the training data, i.e. its value is set to one if the pitch n is active in the frame t, and zero otherwise. The rest of the MEI model parameters (Table 1) are initialized to positive random values. All the parameters are then updated iteratively, as described in Algorithm 1, until the algorithm converges. The updating equations are as follows and they are obtained by applying Equation 7 to each of the model parameters.

$$g_{n,j}(t) \leftarrow g_{n,j}(t) \frac{\sum_{f,m,i} w_{i,n,j} v_{i,m,j} h_j(f) X(t,f) \hat{X}(t,f)^{\beta-2} G(f - m f_0(n))}{\sum_{f,m,i} w_{i,n,j} v_{i,m,j} h_j(f) \hat{X}(t,f)^{\beta-1} G(f - m f_0(n))}, \quad (8)$$

$$h_j(f) \leftarrow h_j(f) \frac{\sum_{t,m,n,i} w_{i,n,j} v_{i,m,j} X(t,f) \hat{X}(t,f)^{\beta-2} G(f - m f_0(n))}{\sum_{t,m,i} w_{i,n,j} v_{i,m,j} \hat{X}(t,f)^{\beta-1} G(f - m f_0(n))}, \quad (9)$$

$$v_{i,m,j} \leftarrow v_{i,m,j} \frac{\sum_{t,f,n} h_j(f) w_{i,n,j} X(t,f) \hat{X}(t,f)^{\beta-2} G(f - m f_0(n))}{\sum_{t,f,n} h_j(f) w_{i,n,j} \hat{X}(t,f)^{\beta-1} G(f - m f_0(n))}, \quad (10)$$

$$w_{i,n,j} \leftarrow w_{i,n,j} \frac{\sum_{t,f,m} h_j(f) v_{i,m,j} X(t,f) \hat{X}(t,f)^{\beta-2} G(f - m f_0(n))}{\sum_{t,f,m,i} h_j(f) v_{i,m,j} \hat{X}(t,f)^{\beta-1} G(f - m f_0(n))}. \quad (11)$$

It is important to remember that $g_{n,j}(t)$ represents the gain for note n at frame t for instrument j. Also $h_j(f)$ is the filter that represents the resonating body of each instrument. Each $v_{i,m,j}$ is an excitation basis vector. There are m excitation basis vectors for each instrument, which are linearly combined with the corresponding weights $w_{i,n,j}$.

Algorithm 1 Training algorithm description

1. Compute $X(t, f)$ from a solo performance for each instrument in the training database.
2. Initialize gains $g_{n,j}(t)$ with the ground-truth pitch transcription and the rest of parameters $h_j(f)$, $v_{i,m,j}$ and $w_{i,n,j}$ with random positive values.
3. Update source-filter $h_j(f)$ following Equation 9.
4. Update excitation basis vectors $v_{i,m,j}$ following Equation 10.
5. Update the weights of the excitation basis vectors $w_{i,n,j}$ following Equation 11.
6. Update gains $g_{n,j}(t)$ following Equation 8.
7. Repeat steps 3-6 until the algorithm converges (or the maximum number of iterations is reached).
8. Compute spectral basis functions $b_{n,j}(f)$ for each instrument j using Equation 5.

Once the MEI parameters are estimated with the NMF framework, the spectral basis functions $b_{n,j}(f)$ for each pitch of the instrument are computed. As a frequency resolution of 1/8 semitones is used, each index from f represents one of the resulting frequency ranges. The training procedure is summarized at Algorithm 1.

Each spectral basis function $b_{n,j}(f)$ required at the factorization stage is computed by the training algorithm. For practical applications, the trained instrument models and the real played instruments have some differences (the trained models are not obtained from the same physical instruments). In this article, we first apply the trained instrument models to perform the separation. We then describe a way to adapt the trained instrument models to the real played instruments.

3.2 Separation with fixed instrument models

This is the basic separation algorithm. Here the NMF factorization framework is composed of two parameters, the gains $g_{n,j}(t)$ and the spectral patterns $b_{n,j}(f)$, as described in Equation 1, but it now includes more than one instrument, specified by index j. Algorithm 2 shows an overview of the separation system with fixed instrument models.

Algorithm 2 Separartion algorithm with fixed instrument models

1. Compute $X(t, f)$ from a the audio mixture to separation.
2. Initialize gains $g_{n,j}(t)$ with the ground-truth pitch transcription and the basis functions $b_{n,j}(f)$ with the trained ones from Algorithm 1.
3. **for** C iterations **do**
4. Update gains $g_{n,j}(t)$ following Equation 8
5. **end for**

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

The MIDI score-provided information is used to initialize the gains $g_{n,j}(t)$ for the NMF factorization. A random positive value is given when the aligned MIDI score indicates that the corresponding pitch indexes of instrument j are active at frame t. The gains associated with non-active pitches of the instrument j at frame t are set to zero. The process to align the score with the mixed signal is explained in Section 2.1 and it follows Duan and Pardo (2011).

Once the gains $g_{n,j}(t)$ are initialized, and using the trained instrument models $b_{n,j}(f)$, an iterative algorithm is run as shown in Algorithm 2. This algorithm iteratively updates $g_{n,j}(t)$ according to Equation 8, while keeping $b_{n,j}(f)$ fixed. Once the gains are estimated, the separated signals are computed as is described in Section 3.5.

Although the NMF factorization framework is intrinsically offline, when spectral patterns $b_{n,j}(f)$ are fixed, the only parameters to be updated are the gains $g_{n,j}(t)$. When using the updating equation (see Equation 8), the gains at frame t just depend on the input signal $X(t, f)$ at frame t and, consequently, the algorithm becomes online. Another interpretation is that the factorization only has to compute the gain at frame t associated to each fixed spectral pattern that minimizes the divergence between the input signal and the reconstructed signal at frame t. In another work (Kim & Park, 2008), this problem is solved using nonnegative least squares which allows one to implement low complexity algorithms but is only suitable for Euclidean distance ($\beta = 2$).

3.3 Adapt instrument models

The pre-learned instrument models are an approximation to the real spectral patterns occurring in the mixture. However, mismatch between training and testing instruments can be large due to the physical differences between the training and testing instruments, the performing style difference of the performers, and the acoustical difference of the recording environments. We propose to adapt the learned models to the instruments in the mixed signal during the factorization process. In this way, the models will fit better with the recorded signal, improving separation results. The adaptation of instrument models with the score alignment information is the main novelty of this work.

To do so, we initialize $g_{n,j}(t)$ from the aligned MIDI score-provided information as in the basic separation method, and also initialize instrument models to the pre-learned ones as described in Section 3.2. The adaptation of the MEI model was also addressed in Carabias-Orti et al. (2011) but without using the score information. In that scenario, the adaptation of instrument models suffers from the interference between overlapping harmonics from different instruments in the polyphonic excerpts. This problem can be handled when the score information is available. Also, the results vary as a function of the parameters updated in the model. In Carabias-Orti et al. (2011), it is shown that results are seriously degraded when adapting the weights $w_{i,n,j}$. This is caused by the huge number of free parameters to be adapted, in contrast to the number required to adapt the excitation basis vectors $v_{i,m,j}$. However, adapting the instrument filter $h_j(f)$ does not degrade the results.

In the proposed separation system, the score alignment information is available. But mistakes at the alignment process have consequences for the updating procedure. The model is sometimes updated with information that does not correspond to the notes played by the instruments. In order to relieve alignment errors affecting the model adaptation, we propose to use fixed weights $w_{i,n,j}$. As reported in Carabias-Orti et al. (2011), the adaptation of weights is very sensitive, due to the large number of parameters to be updated. In the example of the clarinet model, the weights have 592 values while excitation basis vectors have only 40 values. Updating of the instrument filter $h_j(f)$ and the excitation basis vectors $v_{i,m,j}$ but not the weights $w_{i,n,j}$ lets us fit the model to the mixture audio while keeping the model robust to alignment errors. Also, preliminary tests showed this approach gets better results than updating all model parameters.

At each iteration of the NMF algorithm, we update the gains $g_{n,j}(t)$, the instrument filters $h_j(f)$ and the excitation basis vectors $v_{i,m,j}$. At each iteration, the gains updating equation is computed using Equation 8, as in Section 3.2. However, the updating of the instrument model parameters $h_j(f)$ and $v_{i,m,j}$ cannot be computed with the same equations as in the training stage. In the training stage, each note is played alone, so there are no overlapped partials. In the separation stage, notes from different instruments are played simultaneously and some of their harmonics can be overlapped. In these cases, the information for adapting the models from the overlapping harmonics is corrupted due to interference. Depending on the relative phase difference between the overlapped partials at each frame, this interference can be constructive or destructive. Consequently, the adaptation of model parameters should be implemented without influence from the partials where multiple instruments overlap.

We use the aligned score to identify time–frequency regions where instruments may have partials that overlap. After initializing the gains with the score information, the estimated signal for instrument j with the trained models can be computed as

$$\hat{X}_j(t, f) = \sum_{n,m,i} g_{n,j}(t) h_j(m f_0(n)) w_{i,n,j} v_{i,m,j} G(f - m f_0(n)). \tag{12}$$

To select the overlapped time–frequency zones for each instrument $\{f', t', j'\}$, an energy estimation per instrument is computed by using Equation 12. When the energy estimation for instrument j' is not predominant at a time–frequency point (t', f'), this point is added to the overlapped time–frequency set $\{f', t', j'\}$. The energy of instrument j' is considered as predominant, at a time–frequency point (t', f'), when the ratio between its own estimated energy $|\hat{X}_{j'}(t, f)|^2$ and the energy of the rest of instruments $\sum_{j=1, j\neq j'}^{J} |\hat{X}_j(t, f)|^2$, is above 10 dB. Otherwise, this is considered as an overlapped

time–frequency point. Therefore, the overlapped time–frequency zones $\{f', t', j'\}$ for instrument j' are those time–frequency points that fulfill the following restriction

$$10\log_{10}\left(\frac{|\hat{X}_{j'}(t,\,f)|^2}{\sum_{j=1,j\neq j'}^{J}|\hat{X}_j^2(t,\,f)|^2}\right) < 10\,\text{dB}.$$

This estimation is made for all instruments.

Once non-overlapped time–frequency zones for all instruments are estimated, model adaptation can be performed. At the training stage $h_j(f)$ and $v_{i,m,j}$ are updated with information from the whole time and frequency axis. Now updating equations (Equation 9 and Equation 10) are computed in different time–frequency zones for each instrument. For instrument j', signal input $X(t,\,f)$ and reconstructed signal $\hat{X}(f,t)$ are set to zero for the overlapped time–frequency regions $\{f', t', j'\}$. In this way, the instrument models are updated without including the information of overlapped partials. All this computation is summarized in Algorithm 3.

Algorithm 3 Offline separation algorithm with instrument model adaptation

1 Compute $X(t,\,f)$ from a the audio mixture to separation.
2 Initialize $g_{n,j}(t)$ with the score-provided information.
3 Initialize $h_j(f)$, $v_{i,m,j}$ and $w_{i,n,j}$ to the trained instrument models parameters from Algorithm 1.
4 Identify the overlapped time–frequency regions $\{f', t', j'\}$ that satisfy $\frac{|\hat{X}_{j'}^2(t,f)|^2}{\sum_{j=1,j\neq j'}^{J}|\hat{X}_j^2(t,f)|^2} < 10$.
5 **for** C iterations **do**
6 Update gains $g_{n,j}(t)$ with Equation 8.
7 Update the filters $h_j(f)$ and the excitation basis vectors $v_{i,m,j}$ with Equation 9 and Equation 10 only using non-overlapped time–frequency regions for each instrument (through setting $X(f,t)$ and $\hat{X}(f,t)$ to zero in overlapped regions).
8 Compute spectral basis functions $b_{n,j}(f)$ for each instrument j using Equation 5.
9 **end for**

Algorithm 3 iteratively updates instrument models during signal factorization to adapt the models to the instruments played in the mixture. This leads to improved separation results, compared to fixed instrument models, as will be shown in Section 4.4. However, this algorithm requires access to the entire signal and cannot be performed in real-time scenarios. In the next section, we will modify the algorithm to make it work in an online fashion.

3.4 Make the adaptation online

The separation process is considered online when, at each frame, the separated sources can be estimated from only the current and previous frames. Consequently, the instrument models used for each frame factorization should be obtained only with the information from the first frame up to the current one. Because of that, instrument models for the initial frame must be the ones derived in a prior training phase (see Section 3.1). As time moves forward, instrument models should be updated with current and previous observations, so the factorization of the new frames can be computed with an improved version of instrument models. Since this approach does not incorporate information from the future, the separation results should be degraded in comparison with the offline approach.

If we access all the previous frames to update the instrument model in the current frame, then the computational complexity will increase over time. We propose to update instrument models only with the spectral information from $t - T_{update}$ to $t-1$ frames. With this approach we maintain the computational cost constant with time. In addition, we only make the update once every T_{update} frames. Here, updating windows of one second are considered for updating the models ($T_{update} = 1$ s). As in Section 3.3, the overlapped time–frequency regions for each instrument are estimated in order to avoid partials with interference from other sound sources. This process is summarized in Algorithm 4.

In the proposed online system, both the alignment and the factorization stages are computed without any future information, so the system generates the output of a frame t after receiving it. We have used $C = 50$ for the number of iterations. This value was selected as reasonable from a pilot study.

This online algorithm has a very low algorithmic latency (half of a frame). This means that it does not use information from the future, it uses only past information. The system has two stages, the alignment process and the separation stage. In Duan and Pardo (2011), it is shown that the computational complexity of the alignment stage takes $O(R+K)$, where R is the number of particles (on the order of 1000) and K is the number of spectral peaks (on the order of 100). The separation stage takes $O(JF)$, where J represents the instruments (on the order of 10) and F represents the frequency values (on the order of 1000).

If the system were implemented efficiently on a suitable platform, the algorithm has the capacity to work in real-time. It is important to assess how far our current implementation is from real-time computation. In our experiments, the proposed online system is implemented in Matlab on a quad-core 3.2 GHz CPU under Linux. It runs about 7.5 times slower than real time.

3.5 From the estimated gains to the separated signals

3.5.1 Ideal Wiener masks

In this paper, the sources $s_j(t)$, $j = 1...J$ that compose the mixed signal $x(t)$ are linearly mixed, so $x(t) = \sum_{j=1}^{J} s_j(t)$. If the power spectral density of source j at TF bin (f,t) is denoted as $|X_j(t,\,f)|^2$, $j = 1...J$, then, each ideally separated source $s_j(t)$ can be estimated from the mixture $x(t)$ using a generalized time–frequency Wiener filter over the Short-Time

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Fourier Transform (STFT) domain as in Fevotte, Bertin, and Durrieu (2009), and Fritsch and Plumbley (2013). The Wiener filter $\alpha_{j'}$ of source j' represents the relative energy contribution of each source with respect to the energy of the mixed signal $x(t)$. The Wiener filter $\alpha_{j'}$ for each time–frequency bin (t, f) is defined as,

$$\alpha_{j'}(t, f) = \frac{|X_{j'}(t, f)|^2}{\sum_j |X_j(t, f)|^2}, \tag{13}$$

where the estimation of magnitude spectrogram per instrument $X_{j'}(t, f)$ is computed following Equation 12. The sum of all the estimated sources power spectrograms $|\hat{X}_{j'}(t, f)|^2$ are the power spectrogram of the mixed signal $|X(f, t)|^2$. Then, to obtain the estimated source magnitude spectrogram $\hat{X}_{j'}(t, f)$ Equation 14 is used.

$$\hat{X}_{j'}(t, f) = \sqrt{\alpha_{j'}(t, f)} \cdot X(t, f). \tag{14}$$

Finally, the estimated source $\hat{s}_{j'}(t)$ is computed by the inverse overlap-add STFT of the estimated magnitude spectrogram $\hat{X}_{j'}(t, f)$ with the phase spectrogram of the input mixture.

3.5.2 Separated signal decomposition

Once the gains are estimated with any of the proposed methods, the estimated signal recomposition is always the same. First of all, the estimated source magnitude spectrogram $\hat{X}_j(t, f)$ is computed as:

$$\hat{X}_j(t, f) = g_{n,j}(t) b_{n,j}(f). \tag{15}$$

Then, Wiener masks are estimated using Equation 13. These estimated Wiener masks are applied to the mixed signal spectrogram $X(f, t)$ following Equation 14. Finally, the estimated source spectrogram $\hat{X}_j(f, t)$ is obtained and the estimated source $\hat{s}_j(t)$ is computed by the inverse overlap-add STFT over $\hat{X}_j(f, t)$.

4. Experiments

4.1 Training and testing data

At the training stage (see Section 3.1), the spectral basis functions are estimated using the RWC musical instrument sound database (Goto, 2004; Goto et al., 2002). Four instruments are studied in the experiments (violin, clarinet, tenor saxophone and bassoon). The training data included isolated notes for each instrument, recorded at each semitone throughout the entire pitch range of the instrument. The spectral basis functions for each instrument are estimated from all the note recordings. Files from the RWC database have different playing styles. Files with a normal playing style and mezzo dynamic level are selected, as in prior literature. Training with different playing styles leads to different models. However, as demonstrated in Carabias-Orti et al. (2011), the selected configuration (normal playing style and mezzo dynamic level) is representative for the different models.

The database proposed in Duan and Pardo (2011) is used for the testing stage. This database consists of 10 J.S. Bach four-part chorales with the corresponding aligned MIDI data. The audio files are approximately 30 s long and are sampled at 44.1 kHz from 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. Individual tracks from the each chorale were mixed to create 60 duets, 40 trios, and 10 quartets, totalling 110 polyphonic music audio performances. The scores were MIDI downloaded from the internet[1]. The ground-truth alignment between MIDI and audio was interpolated from annotated beat times of the audio. The annotated beats were verified by a musician through playing back the audio together with these beats as explained in Duan and Pardo (2011).

4.2 Experimental set-up

4.2.1 Time-frequency representation

Many NMF-based signal processing applications adopt frequency logarithmic discretization. For example, uniformly spaced subbands on the Equivalent Rectangular Bandwidth (ERB) scale are assumed in Bertin, Badeau and Vincent (2010), and Vincent and Ono (2010). When using instrument models with harmonic restrictions, the reconstructed signal is computed with a term derived from the window transform $G(f - mf_0)$ translated to the pitch-dependent frequency mf_0. This term appears in Equation 5 when the MEI model is used. In this scenario, a frequency resolution related to the pitch resolution is recommended to facilitate signal computation. Additionally, the training database and the ground-truth score information are composed of notes that are separated by one semitone in pitch. Here, we use a frequency resolution of 1/8 of a semitone. In this work, we implement a time–frequency representation by integrating the STFT bins corresponding to the same 1/8 semitone interval. When computing the separation Wiener masks, the same mask value is applied to all the frequency bins belonging to the same 1/8 of semitone interval.

4.2.2 Model parameters

The frame size and the hop size for the STFT are set to 128 and 32 ms respectively. Also, $C = 50$ iterations for the NMF-based algorithms is used.

The MEI model is computed with the following parameters: (1) 20 harmonics per spectral basis function for the harmonic constraint models ($M = 20$); (2) $I = 2$ excitation basis vectors; and (3) $J = 4$ instruments, the same as in the test database.

In relation to the pitch resolution, in the training stage a pitch resolution of a semitone (the same as the training database) is used. In the separation stage, the learned basis functions $b_{n,j}(f)$ are adapted to a 1/8 semitone resolution in pitch

[1] http://www.jsbchorales.net/index.shtml

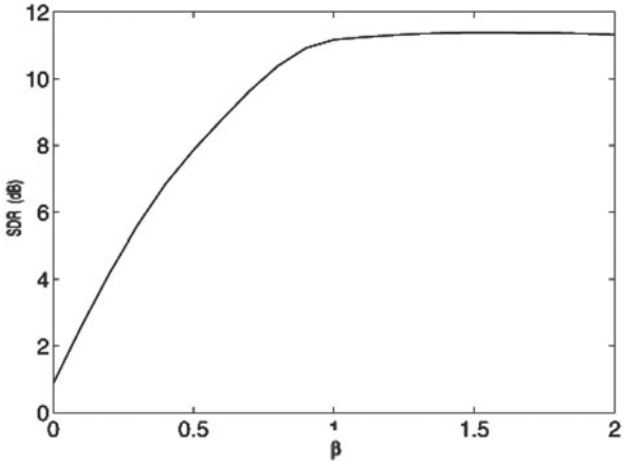

Fig. 1. Performance of a polyphony level 2 separation with different values of β. Higher SDR values are better.

by replicating the function for each semitone 8 times. Real instruments produce pitches not only at idealized semitones, due to pitch variations such as vibrato. With a 1/8 semitone resolution in pitch, we can better capture the pitch variation of real instruments.

The use of the β-divergence distortion in the NMF framework lets us set the parameter β to the value that obtains the best results. To find the optimum value of this parameter, we performed source separation over 60 duets (all the duets from the Bach chorales database), varying the value of β in the range [0, 2] with a step size of 0.1. We repeated this 10 times for each value of β. Figure 1 shows that the optimal value of β is around $\beta = 1.5$. A paired t-test showed a statistically significant difference ($p < 10^{-3}$) for β values under 1.1 and above 1.7 with respect $\beta = 1.5$. However, no statistically significant differences are found in the interval of $\beta = [1.2, 1.6]$. Therefore, we used $\beta = 1.5$ for the remainder of our experiments. These results are similar to the ones from Fitzgerald, Cranitch and Coyle (2008). Despite the fact that different test databases have been used, the ranges of β with better SDR values are similar. In Fitzgerald et al. (2008), this range is set between 0.8 and 1.4.

Although lower values of β have been proven to be very suitable in other signal processing applications (Bertin et al., 2010; Fevotte et al., 2009), here the performance is very poor. The reason for this is related to the harmonicity constraint imposed here. The musical instruments are not perfectly harmonic and they generate low energy values outside the neighbourhood of the harmonics that cannot be modelled with the used spectral basis functions. The same occurs for the background noise present in the signal. Low energy values and background noise do not modify the distortion measure when using high values of β because they more heavily rely on the largest values, but in the case of β values close to 0, these differences primarily are represented at the divergence measure because of the scale invariance in the case of $\beta = 0$ (Carabias-Orti et al., 2011).

4.2.3 Audio separation metrics

For an objective evaluation of the source separation performance of the proposed method, we use the metrics implemented in Vincent, Gribonval and Fevotte (2006) (BSS EVAL Toolbox 2.1). These metrics are commonly accepted by the research community in source separation, and therefore facilitate a fair evaluation of the method. Each separated signal is assumed to produce a distortion model that can be expressed as follows,

$$\hat{s}_j(t) - s_j(t) = e_j^{target}(t) + e_j^{interf}(t) + e_j^{artif}(t), \quad (16)$$

where $\hat{s}_j$ is the estimated source signal for instrument j, s_j is the original signal of the instrument j, e^{target} is the error term associated with the target distortion component, e^{interf} is the error term due to interference of the other sources and e^{artif} is the error term attributed to the numerical artifacts of the separation algorithm. The metrics for each separated signal are the *Source to Distortion Ratio* (SDR), the *Source to Interference Ratio* (SIR), and the *Source to Artifacts Ratio* (SAR).

4.3 Algorithms for comparison

We compare different configurations of the proposed method and a baseline score-informed source separation method proposed in Duan and Pardo (2011), denoted as *Soundprism*. It separates sources using harmonic masking where the energy of overlapping harmonics are distributed according to the harmonic indices of the sources. It is an online algorithm but no instrument models are used.

The proposed method has three configurations. *Proposed fixed* denotes the online version of the proposed method using fixed instrument models (Section 3.2). *Proposed adaptive offline* denotes the offline version of the proposed method with adaptive instrument models (Section 3.3), and *Proposed adaptive online* denotes its online version (Section 3.4).

We also compare with *Oracle*, the theoretically best source separation method based on time–frequency masking methods and the analysis filter bank used on the proposed separation system. Its calculation requires the isolated sound sources. The mixed signals are filtered with the analysis filter bank of 1/8 of semitone frequency resolution. After that, the isolated sources, that are also filtered by the analysis filter bank, are used to obtain the ideal Wienner masks, which are the best mask that can be obtained with the given frequency resolution. Then, these masks are applied to the mixed signal and the oracle separated signal are obtained. This process gives the best possible separation with the system set-up. It sets an upper bound of all the configurations of the proposed method.

4.4 Results

4.4.1 Working with ground-truth pitches

The proposed source separation method is intended to work in the score-informed scenarios. However, errors in the

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

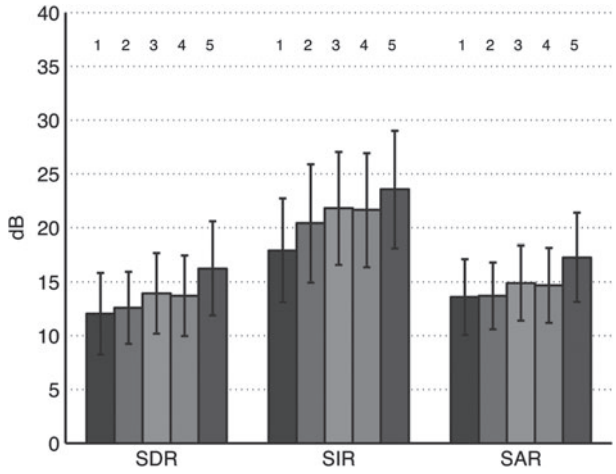

Fig. 2. Source separation results on the 60 duets using ground-truth pitch information. Each bar shows the average of 120 measurements on the 120 separated tracks. The vertical line around the top of each bar shows the plus and minus standard deviation. The five methods are (1) Soundprism, (2) Proposed fixed (Section 3.2), (3) Proposed adaptive offline (Section 3.3), (4) Proposed adaptive online (Section 3.4), and (5) Oracle, labelled above the bars. Higher values are better.

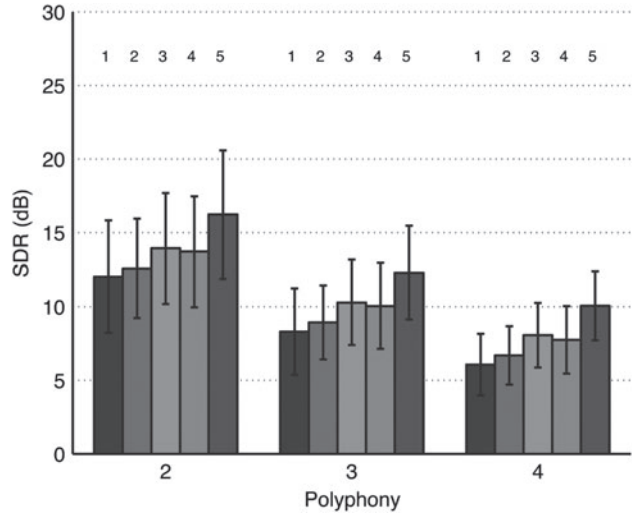

Fig. 3. Source separation results versus polyphony, calculated using the ground-truth pitch information. Each bar is the average of 120 measurements for duets and triples, and of 40 measures for quartets, where one measurement is calculated for each separated track. The vertical line around the top of each bar shows the plus and minus standard deviation. The five methods are (1) Soundprism, (2) Proposed fixed (Section 3.2), (3) Proposed adaptive offline (Section 3.3), (4) Proposed adaptive online (Section 3.4), and (5) Oracle, labelled above the bars. Higher values are better.

audio-score alignment stage may also affect source separation. In order to separate out this effect and focus on the separation algorithms, we first test the proposed method with ground-truth pitches. The ground-truth pitches were obtained by running YIN (de Cheveigné & Kawahara, 2002) on the isolated source signals before mixing, followed by necessary manual corrections.

Figure 2 shows the comparison results on the 60 duets. We can see all methods including the Oracle (Bar 5) show a pretty large standard deviation. This is due to the variations of difficulty in separating different musical instruments. Compared to the baseline Soundprism method (Bar 1), the proposed method of all three configurations (Bars 2, 3 and 4) improves significantly on SDR and SIR. A one-sided paired t-test is performed to evaluate the significance. Results show that the improvements are all statistically significant ($p < 10^{-6}$), compared to the baseline Soundprism. In terms of SAR, the proposed method with fixed instrument models (Bar 2) is not statistically better than Soundprism (Bar 1), but the proposed method with adaptive instrument models (Bar 3) is statistically significantly better than Soundprism (Bar 1). In fact, the introduction of adaptive instrument models (Bar 3) improves all three metrics, compared to using fixed instrument models (Bar 2). This improvement is statistically significant ($p < 10^{-5}$). The online algorithm with adaptive instrument models (Bar 4) has slightly worse performance on all three metrics than its offline version (Bar 3), but it is still significantly better than the offline algorithm using fixed instrument models (Bar 2) and the online algorithm without instrument models (i.e. Soundprism, Bar 1). Compared to Oracle results (Bar 5), we can see the proposed online adaptive algorithm (Bar 4) achieves about 2.5 dB lower SDR, which leaves room for the algorithm to improve.

Figure 3 shows the comparison results on recordings of different polyphonies. Only SDR is shown, as the trends for SIR and SAR are similar. With increasing polyphony, the performance of all methods, including the Oracle, decreases significantly. Similar to the results for duets in Figure 2, all three configurations (Bars 2, 3 and 4) of the proposed method improve on Soundprism (Bar 1). This improvement is statistically significant, as confirmed by a one-sided paired t-test ($p < 10^{-4}$). In addition, the improvement from fixed instrument models (Bar 2) to adaptive instrument models (Bars 3 and 4) is also statistically significant ($p < 10^{-8}$).

4.4.2 Working with audio-score alignment

In this section, we compare source separation methods taking audio-score alignment results (i.e. the score pitches) as inputs. This evaluates the proposed method in realistic situations. Figure 4 shows the results on duets. Similar to Figure 2, the proposed method using adaptive instrument models (Bar 3 and 4) significantly outperforms Soundprism (Bar 1) in SDR and SIR, which is confirmed by an one-sided paired t-test ($p < 10^{-3}$). The proposed method using fixed instrument models (Bar 2) significantly outperforms Soundprism (Bar 1) in SIR ($p = 4.1 \times 10^{-5}$), but not in SDR ($p = 0.16$). This again shows the benefit of using (adaptive) instrument models for source separation. The improvement from using fixed instrument models (Bar 2) to adaptive instrument models (Bar 3 and 4) is again statistically significant on all three metrics ($p < 10^{-4}$), for both offline and online algorithms, even

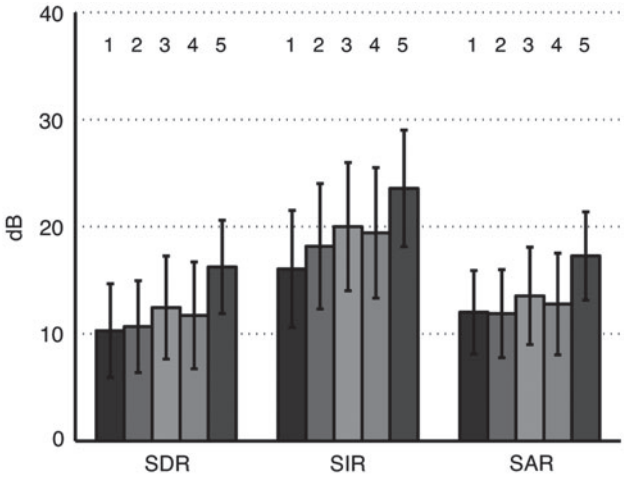

Fig. 4. Source separation results on the 60 duets using the aligned musical score information. Each bar shows the average of 120 measurements on the 120 separated tracks. The vertical line around the top of each bar shows the plus and minus standard deviation. The five methods are (1) Soundprism, (2) Proposed fixed (Section 3.2), (3) Proposed adaptive offline (Section 3.3), (4) Proposed adaptive online (Section 3.4), and (5) Oracle, labelled above the bars.

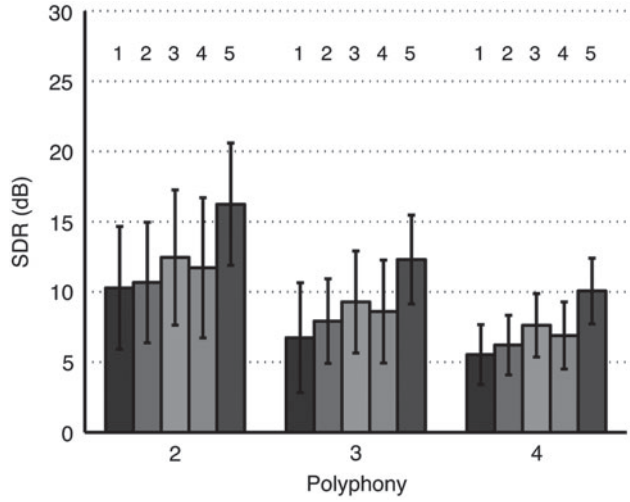

Fig. 5. Source separation results versus polyphony, calculated using the alignment information. Each bar is the average of 120 measurements for duets and triples, and of 40 measures for quartets, where one measurement is calculated for each separated track. The vertical line around the top of each bar shows the plus and minus standard deviation. The five methods are (1) Soundprism, (2) Proposed fixed (Section 3.2), (3) Proposed adaptive offline (Section 3.3), (4) Proposed adaptive online (Section 3.4), and (5) Oracle, labelled above the bars.

though the online algorithm (Bar 4) drops the performance a little from its offline version (Bar 3).

Figure 5 further shows the results on recordings of different polyphonies. Similar to Figure 3, only SDR values are shown as the trends for SIR and SAR are similar. Again, the proposed method with all three configurations (Bars 2, 3 and 4) significantly outperforms Soundprism (Bar 1) for all polyphonies, which is confirmed by a one-sided paired t-test ($p < 10^{-4}$), with the exception of duets. The improvement from fixed instrument models (Bar 2) to adaptive models (Bar 3 and 4) is also statistically significant for all polyphonies ($p < 10^{-5}$). These results shows the advantage of using (adaptive) instrument models over not using instrument models in score-informed source separation, and also shows that the proposed online algorithm is able to retain this advantage.

Compared with the results using ground-truth pitch information in Figure 3, there are two additional interesting observations. First, the average SDR of all methods in Figure 5 except Oracle decreases and the standard deviation increases. This is because of the audio-score alignment errors. Second, with the increase of polyphony, the degradations are less significant for almost all methods. This can be explained by the performance of the audio-score alignment. On this dataset, the alignment was better on pieces with higher polyphony (Duan & Pardo, 2011).

5. Conclusions and discussions

In this work, a score-informed source separation model is proposed. It uses instrument models that describe the spectral behaviour of each instrument. Different configurations have been tested and compared to a state-of-the-art method (*Soundprism*) as baseline and the *Oracle* separation.

Unlike existing systems, our system uses parametrized instrument models learned from non-synthetic performances. The models are adapted to the real instrument from the input signal while computing the separation. This adaptation is done by considering only the non-overlapped partials. Information about which partials overlap is obtained from a performance-aligned score. Since one version of our source separation with adaptive instrument models can be performed without any future information, it can be considered an online algorithm.

The system has been tested over a state-of-the-art musical database and compared with a state-of-the-art system that does not use instrument models. The experiments show that the use of instrument models improves the separation results. Furthermore, the separation performance with adaptive instrument models is better than that with fixed models. The online algorithm results are nearly as good as the offline ones and the same occurs when using online alignment information. The difference between online and offline performance is reduced as the polyphony grows because of the better performance of the alignment stage as polyphony increases.

Despite using only one basis function per note in the MEI model, the experimental section of this work shows that it achieves promising results when separating musical performances with a moderate dynamic variation. If a wide dynamic variation were presented, more than one basis function would be needed for representing each note. Some instruments have a particular spectral shape when the amplitude level changes. For instance, the clarinet spectral shape relies on non-linear mechanisms. That is why its spectrum depends upon the gain. Here, it is supposed that most of the time, the instrument

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

amplitude levels are in a medium range, so that the spectral shape is stable. For future work, it could be interesting to model the dependence of spectral shape on amplitude as has been performed in Dannenberg and Derenyi (1998), Hu (2004), and Horner and Beauchamp (2008).

In this work, the instrument models are updated with non-overlapped partials, while the overlapped partials are only modulated by the separation masks. In future work, a mixed separation framework can be developed which separates the non-overlapped partials with the mask procedure and the overlapped ones with sound synthesis in the time domain after estimating both amplitude and phase parameters.

Acknowledgements

We thank the reviewers for their thorough and constructive comments which helped improve the paper significantly.

Funding

This work was supported by the Andalusian Business, Science and Innovation Council under project P2010-TIC-6762 and (FEDER) the Spanish Ministry of Economy and Competitiveness under Project TEC2012-38142-C04-03.

References

Babaie-zadeh, M., & Jutten, C. (2006). Semi-blind approaches for source separation and independent component analysis. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(3), 538–549.

Badeau, R., Emiya, V., & David, B. (2009). Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra. *Paper presented at the International Conference on Acoustics, Speech, Signal Processing (ICASSP)*. Taipei, Taiwan.

Bertin, N., Badeau, R., & Vincent, E. (2010). Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(3), 538-549.

Bryan, D.D., Lee, & Seung, H.S. (2000). Algorithms for non-negative matrix factorization. *Proceedings of the Neural Information Processing Systems (NIPS), Denver, CO, USA,* 556–562.

Carabias-Orti, J.J., Virtanen, T., Vera-Candeas, P., Ruiz-Reyes, N., & Cañadas-Quesada, F.J. (2011). Musical instrument sound multi-excitation model for non-negative spectrogram factorization. *IEEE Journal of Selected Topics in Signal Processing, 5*(6), 1144-1158.

Chordia, P., & Rae, A. (2009). Using source separation to improve tempo detection. *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*. (pp. 183–188). Canada: International Society for Music Information Retrieval.

Cichocki, A., & Zdunek, R. (2007). Regularized alternating least squares algorithms for non-negative matrix/tensor factorization. *Advances in Neural Networks - ISNN 2007*. (pp. 793–802). Berlin, Heidelberg: Springer

Comon, P., & Jutten, C. (2010). *Handbook of blind source separation: Independent component analysis and applications*. New York: Academic Press.

Cont, A. (2006). Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (pp. 245–248). Piscataway, NJ: IEEE

Cont, A. (2010). A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analasis and Machine Intelligence, 32*(6), 974–987.

Dannenberg, R.B., & Derenyi, I. (1998). Combining instrument and performance models for high-quality music synthesis. *Journal of New Music Research, 27*(3), 211-238.

de Cheveigné, A., & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, shape, 111*, 1917–1930.

Dixon, S., & Widmer, G. (2005). MATCH: A music alignment tool chest. *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*. (pp. 492–497). London: Queen Mary, University of London.

Duan, Z., Mysore, G., & Smaragdis, P. (2012). Online PLCA for real-time semi-supervised source separation. *Paper presented at the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2012)* (pp. 34–41). Tel-Aviv, Israel.

Duan, Z., & Pardo, B. (2011). Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing, 5*(6), 1205–1215. doi:10.1109/JSTSP.2011.2159701.

Duan, Z., Pardo, B., & Zhang, Z. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transacctions on Audio, Speech & Language Processing, 18*(8), 2121–2133.

Ewert, S., & Muller, M. (2012). Using score-informed constraints for NMF-based source separation. *Proceedings of the IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan*. Piscataway, NJ: IEEE. (pp. 129–132)

Fevotte, C., Bertin, N., & Durrieu, J.L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence. *With application to music analysis, Neural Computation, 21*(3), 793–830.

Fevotte, C., & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation, 23*(9), 2421–2456.

Fitzgerald, D., Cranitch, M., & Coyle, E. (2008). On the use of the beta divergence for musical source separation. *Paper presented at the Irish Signals and Systems Conference*, Ireland.

Fritsch, J., & Plumbley, M.D. (2013). Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada. Piscataway, NJ: IEEE.

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Downloaded by [Biblioteca, Universitat Pompeu Fabra] at 04:23 20 May 2015

Gainza, M., & Coyle, E. (2007). Automating ornamentation transcription. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007*. (Vol. 1, pp. I-69–I-72) Piscataway, NJ: IEEE.

Ganseman, J., Scheunders, P., Mysore, G., & Abel, J. (2010). Evaluation of a score-informed source separation system. *11th International Society for Music Information Retrieval Conf (ISMIR 2010)* (pp. 219–224). Canada: International Society for Music Information Retrieval.

Gillis, N., & Luce, R. (2014). Robust near-separable nonnegative matrix factorization using linear optimization. *Journal of Machine Learning Research, 15*, 1249–1280.

Goto, M. (2004). Development of the RWC music database. *Proceedings of the 18th International Congress on Acoustics, (ICA)* (pp. I-553–556, Invited Paper)

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC music database: Popular, classical, and jazz music databases. *Proceedings of the 3rd International Society for Music Information Retrieval Conference (ISMIR), Paris, France* (pp. 287–288). Paris: IRCAM - Centre Pompidou.

Heittola, T., Klapuri, A., & Virtanen, T. (2009). Musical instrument recognition in polyphonic audio using source-filter model for sound separation. *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR), Kobe, Japan* (pp. 327–332). Canada: International Society for Music Information Retrieval.

Hennequin, R., David, B., & Badeau, R. (2011). Score informed audio source separation using a parametric model of non-negative spectrogram. *Paper presented at ICASSP 2011*. Prague, Czech Republic.

Horner, A.B., & Beauchamp, J.W. (2008). *Instrument modeling and synthesis. Handbook of signal processing in acoustics*. New York: Springer. (Vol. 1, pp. 375–397)

Hu, N. (2004). Automatic construction of synthetic musical instruments and performers (PhD thesis), University of California., Berkeley, USA.

Hyvarinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks, 13*, 411–430.

Joder, C., Weninger, F., Eyben, F., Virette, D., & Schuller, B. (2012). Real-time speech separation by semi-supervised nonnegative matrix factorization. *Paper presented at the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)* (pp. 322–329), Tel-Aviv, Israel.

Kim, J., & Park, H. (2008). Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis & Applications, 30*, 713–730.

Klapuri, A., Virtanen, T., & Heittola, T. (2010). Sound source separation in monaural music signals using excitation-filter model and em algorithm. *Paper presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Dallas, TX, USA.

Lee, D.D., & Seung, H.S. (2001). Algorithms for nonnegative matrix factorization. In *Neural Information Processing Systems* (pp. 556–562), Cambridge, MA: MIT Press.

Rabiner, R., & Schafer, R.W. (1978). *Digital processing of speech signals*. Upper Saddle River, NJ: Prentice Hall.

Rodriguez-Serrano, F.J., Carabias-Orti, J.J., Vera-Candeas, P., Canadas-Quesada, F.J., & Ruiz-Reyes, N. (2013). Monophonic constrained non-negative sparse coding using instrument models for audio separation and transcription of monophonic source-based polyphonic mixtures. *Multimedia Tools and Applications, 72*(1), 925–949.

Simon, L.S.R., & Vincent, E. (2012). A general framework for online audio source separation. Latent Variable Analysis and Source Separation. *Paper presented at the 10th International Conference on (LVA/ICA 2012)* (pp. 364–371). Tel-Aviv, Israel:

Simsekli, U., & Cemgil, A.T. (2012). Score guided musical source separation using Generalized Coupled Tensor Factorization. *Paper presented at the 20th European Signal Processing Conf (EUSIPCO), 2012*. (pp. 2639–2643), Bucharest, Romania.

Välimäki, V., Pakarinen, J., Erkut, C., & Karjalainen, M. (2006). Discrete-time modeling of musical instruments. *Reports on Progress in Physics, 39*(1), doi:10.1088/0034-4885/69/1/R01.

Vincent, E., Bertin, N., & Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(3), 528–537.

Vincent, E., Gribonval, R., & Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing, 14*(4), 1462-1469.

Vincent, E., & Ono, N. (2010). *Music source separation and its applications to MIR*. Tutorial presented at The International Society for Music Information Retrieval Conf, (ISMIR); Utrecht, Netherlands.

Virtanen, T., & Klapuri, A. (2006). Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. *Paper presented at the Neural Information Processing Systems Workshop in Advances in Models for Acoustic*, Vancouver, Canada.

Viste, H., & Evangelista, G. (2001). Sound source separation: Preprocessing for hearing aids and structured audio coding. *Paper presented at the COST G-6 Conf on Digital Audio Effects (DAFX)*, Limerick, Ireland.

Zibulevsky, M., Kisilev, P., Zeevi, Y.Y., & Pearlmutter, B. (2002). Blind source separation multinode sparse representation. *Paper presented at the Neural Information Processing Systems Workshop*, Vancouver, Canada.