Professionally-produced music separation guided by covers

Timothée Gerber, Martin Dutasta, Laurent Girin, Cédric Févotte

PROFESSIONALLY-PRODUCED MUSIC SEPARATION GUIDED BY COVERS

Timothée Gerber, Martin Dutasta, Laurent Girin
Grenoble-INP, GIPSA-lab
firstname.lastname@gipsa-lab.grenoble-inp.fr

Cédric Févotte
TELECOM ParisTech, CNRS LTCI
cedric.fevotte@telecom-paristech.fr

ABSTRACT

This paper addresses the problem of demixing professionally produced music, i.e., recovering the musical source signals that compose a (2-channel stereo) commercial mix signal. Inspired by previous studies using MIDI synthesized or hummed signals as external references, we propose to use the multitrack signals of a cover interpretation to guide the separation process with a relevant initialization. This process is carried out within the framework of the multichannel convolutive NMF model and associated EM/MU estimation algorithms. Although subject to the limitations of the convolutive assumption, our experiments confirm the potential of using multitrack cover signals for source separation of commercial music.

1. INTRODUCTION

In this paper, we address the problem of source separation within the framework of professionally-produced (2-channel stereo) music signals. This task consists of recovering the individual signals produced by the different instruments and voices that compose the mix signal. This would offer new perspectives for music active listening, editing and post-production from usual stereo formats (e.g., 5.1 upmixing), whereas those features are currently roughly limited to multitrack formats, in which a very limited number of original commercial songs are distributed.

Demixing professionally produced music (PPM) is particularly difficult for several reasons [11, 12, 17]. Firstly, the mix signals are generally underdetermined, i.e., there are more sources than mix channels. Secondly, some sources do not follow the point source assumption that is often implicit in the (convolutive) source separation models of the signal processing literature. Also, some sources can be panned in the same direction, convolved with large reverberation, or processed with artificial audio effects that are more or less easy to take into account in a separation framework. PPM separation is thus an ill-posed problem and separation methods have evolved from blind to *informed* source separation (ISS), i.e., methods that exploit some "grounded" additional information on the source/mix signals and mix process. For example, the methods in [1,4,5,8,20] exploit the musical score of the instrument to extract sources, either directly or through MIDI signal synthesis. In user-guided approaches, the listener can assist the separation process in different ways, e.g., by humming the source to be extracted [16], or by providing information on the sources direction [19] or temporal activity [12]. An extreme form of ISS can be found in [6, 9, 10, 14, 15] and in the Spatial Audio Object Coding (SAOC) technology recently standardized by MPEG [3]: here, the source signals themselves are used for separation, which makes sense only in a coder-decoder configuration.

In the present paper, we remain in the usual configuration where the original multitrack signals are not available, although we keep the latter spirit of using *source signals* to help the demixing process: we propose to use *cover multitrack signals* for this task. This idea is settled on several facts. Firstly, a cover song can be quite different from the original for the sake of artistic challenge. But very interestingly, for some applications/markets a cover song is on the contrary intended to be as close as possible to the original song: instruments composition and color, song structure (chorus, verses, solos), and artists interpretation (including the voices) are then closely fitted to the original source signals, hence having a potential for source separation of original mixes. Remarkably, it happens that multitracks of such "mimic" covers are relatively easy to find on the market for a large set of famous pop songs. In fact, they are much easier to obtain than original multitracks. This is because the music industry is very reluctant to release original works while it authorizes the licensed production of mimic multitracks on a large scale. In the present study, we will use such multitracks provided by iKlax Media which is a partner of the DReaM project. [1] iKlax Media produces software solutions for music active listening and has licensed the exploitation of a very large set of cover multitracks of popular songs. Therefore, this work involves a sizeable artistic and commercial stake. Note that similar material can be obtained from several other companies.

We set the cover-informed source separation principle within the currently very popular framework of separation methods based on a local time-frequency (TF) complex Gaussian model combined with a non-negative matrix factorization (NMF) model for the source variances [7,11,13].

[1] This research is partly funded by the French National Research Agency (ANR) – Grant CONTINT 09-CORD-006.

Iterative NMF algorithms for source modeling and separation have shown to be very sensitive to initialization. We turn this weakness into strength within the following two-step process in the same spirit as the work carried out on signals synthesized from MIDI scores in, e.g., [8] or by humming in [16]. First, source-wise NMF modeling is applied on the cover multitrack, and the result is assumed to be a suitable initialization of the NMF parameters of the original sources (that were used to produce the commercial mix signal). Starting from those initial values, the NMF process is then refined by applying to the mix the convolutive multichannel NMF model of [11]. This latter model provides both refined estimation of the source-within-mix (aka source images) NMF parameters and source separation using Wiener filters built from those parameters.

The paper is organized as follows. In Sections 2 and 3, we respectively present the models and method employed. In Sections 4 and 5, we present the experiments we conducted to assess the proposed method, and in Section 6, we address some general perspectives.

2. FRAMEWORK: THE CONVOLUTIVE MULTICHANNEL NMF MODEL

2.1 Mixing Model

Following the framework of [11], the PPM multichannel mix signal $x(t)$ is modeled as a convolutive noisy mixture of J source signals $s_j(t)$. Using the short-time Fourier transform (STFT), the mix signal is approximated in the TF domain as:

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{1}$$

where $\mathbf{x}_{fn} = [x_{1,fn}, \ldots, x_{I,fn}]^T$ is the vector of complex-valued STFT coefficients of the mix signal, $\mathbf{s}_{fn} = [s_{1,fn}, \ldots, s_{J,fn}]^T$ is the vector of complex-valued STFT coefficients of the sources, $\mathbf{b}_{fn} = [b_{1,fn}, \ldots, b_{I,fn}]^T$ is a zero-mean Gaussian residual noise, $\mathbf{A}_f = [\mathbf{a}_{1,f}, \ldots, \mathbf{a}_{J,f}]$ is the frequency-dependent mixing matrix of size $I \times J$ ($\mathbf{a}_{j,f}$ is the mixing vector for source j), $f \in [0, F-1]$ is the frequency bin index and $n \in [0, N-1]$ is the time frame index. This approach implies standard narrowband assumption (i.e., the time-domain mixing filters are shorter than the STFT window size).

2.2 Source model

Each source $s_{j,fn}$ is modeled as the sum of K_j latent components $c_{k,fn}$, $k \in \mathcal{K}_j$, i.e.,

$$s_{j,fn} = \sum_{k \in \mathcal{K}_j} c_{k,fn}, \tag{2}$$

where $\{\mathcal{K}_j\}_j$ is a non-trivial partition of $\{1, \ldots, K\}$, $K \geq J$ (K_j is thus the cardinal of $\mathcal{K}_j$). Each component $c_{k,fn}$ is assumed to follow a zero-mean proper complex Gaussian distribution of variance $w_{fk}h_{kn}$, where $w_{fk}, h_{kn} \in \mathbb{R}^+$, i.e., $c_{k,fn} \sim \mathcal{N}_c(0, w_{fk}h_{kn})$. The components are assumed to be mutually independent and individually independent across frequency and time, so that we have:

$$s_{j,fn} \sim \mathcal{N}_c(0, \sum_{k \in \mathcal{K}_j} w_{fk}h_{kn}). \tag{3}$$

This source model corresponds to the popular non-negative matrix factorization (NMF) model as applied to the source power spectrogram $|\mathbf{S}_j|^2 = \{|s_{j,fn}|^2\}_{fn}$:

$$|\mathbf{S}_j|^2 \simeq \mathbf{W}_j \mathbf{H}_j, \tag{4}$$

with non-negative matrices $\mathbf{W}_j = \{w_{fk}\}_{f,k \in \mathcal{K}_j}$ of size $F \times K_j$ and $\mathbf{H}_j = \{h_{kn}\}_{k \in \mathcal{K}_j, n}$ of size $K_j \times N$. The columns of $\mathbf{W}_j$ are generally referred to as *spectral pattern vectors*, and the rows of $\mathbf{H}_j$ are referred to as *temporal activation vectors*. NMF is largely used in audio source separation since it appropriately models a large range of musical sounds by providing harmonic patterns as well as non-harmonic ones (e.g., subband noise).

2.3 Parameter estimation and source separation

In the source modeling context, the NMF parameters of a given source signal can be obtained from the observation of its power spectrogram using Expectation-Maximization (EM) iterative algorithms [7]. In [11], this has been generalized to the joint estimation of the J sets of NMF source parameters and $I \times J \times F$ mixing filters parameters from the observation of the mix signal power spectrogram. More precisely, two algorithms were proposed in [11]. An EM algorithm consists of maximizing the exact joint likelihood of the multichannel data, whereas a multiplicative updates (MU) algorithm, maximizes the sum of individual channel log-likelihood. If the former better exploits the inter-channel dependencies and gives better separation results, [2] the latter has a lower computation cost. Those algorithms will not be described in the present paper, the reader is referred to [11] for technical details.

Once all the parameters are estimated, the source signals (or their spatial images $\mathbf{y}_{j,fn} = \mathbf{a}_{j,f} s_{j,fn}$) are estimated using spatial Wiener filtering of the mix signal:

$$\hat{\mathbf{s}}_{fn} = \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H \mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \mathbf{x}_{fn}, \tag{5}$$

where $\mathbf{\Sigma}_{\mathbf{s},fn}$ is the (estimated) covariance matrix of the source signals, and $\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f}$ is the (estimated) covariance matrix of the mix signal.

3. PROPOSED COVER-INFORMED SEPARATION TECHNIQUE

3.1 Cover-based initialization

It is well-known that NMF decomposition algorithms are highly dependent on the initialization. In fact, the NMF model does not guarantee the convergence to a global minimum but only to a local minimum of the cost function, making a suitable initialization crucial for the separation performance. In the present study, we have at our disposal

[2] When point source and convolutive mixing assumptions are verified.

the 2-channel stereo multitrack cover of each song to separate, and the basic principle is to use the cover source tracks to provide relevant initialization for the joint multichannel decomposition. Therefore, the NMF algorithms mentioned in Section 2 are applied on PPM within the following configuration. A first multichannel NMF decomposition is run on each stereo source of the cover multitrack (with random initialization). Thus, we obtain a modeled version of each cover source signal in the form of three matrices per source: $\mathbf{W}_j^{cover}$, $\mathbf{H}_j^{cover}$ and $\mathbf{A}_j^{cover} = \{a_{ij,f}^{cover}\}_{i\in[1,2],f}$. The results are ordered according to:

$$\mathbf{W}_{init}^{mix} = [\mathbf{W}_1^{cover} \dots \mathbf{W}_J^{cover}] \tag{6}$$

$$\mathbf{H}_{init}^{mix} = \begin{bmatrix} \mathbf{H}_1^{cover} \\ \vdots \\ \mathbf{H}_J^{cover} \end{bmatrix} \tag{7}$$

$$\mathbf{A}_{init}^{mix} = [\mathbf{A}_1^{cover} \dots \mathbf{A}_J^{cover}] \tag{8}$$

Then, (6), (7), and (8) are used as an initialization for a second convolutive stereo NMF decomposition run on the mix signal as in [11]. During this second phase, the spectral pattern vectors and time activation vectors learned from the cover source tracks are expected to evolve to match the ones corresponding to the signals used to produce the commercial mix, while the resulting mixing vectors are expected to fairly model the mix process.

3.2 Pre-processing: time alignment of the cover tracks

One main difference between two versions of the same music piece is often the temporal misalignment due to both tempo variation (global misalignment) and musical interpretation (local misalignments). In a general manner, time misalignment can corrupt the separation performances if the spectral pattern vectors used for initialization are not aligned with the spectral patterns of the sources within the mix. In the present framework, this problem is expected to be limited by the intrinsic automatic matching of temporal activity vectors within the multichannel NMF decomposition algorithm. However, the better the initial alignment, the better the initialization process and thus expected final result. Therefore, we limit this problem by resynchronizing the cover tracks with the mix signal, in the same spirit as the MIDI score-to-audio alignment of [5] or the Dynamic Time Warping (DTW) applied on synthesized signals in [8]. In the present study, this task is performed at quarter-note accuracy using the Beat Detective tool from the professional audio editing software Avid ProTools®. This step allows minimizing synchronization error down to less than a few TF frames, which is in most cases below the synchronization error limit of 200 ms observed in [5]. In-depth study of desynchronization on source separation is kept for future works.

3.3 Exploiting the temporal structure of source signals

In order to further improve the results, we follow a user-guided approach as in [12]. The coefficients of matrix $\mathbf{H}$ are zeroed when the source is not active in the mix, exploiting audio markers of silence zones in the cover source tracks. As there still may be some residual misalignment between the commercial song and the cover after the pre-processing, we relax these constraints to 3 frames before and after the active zone. When using the MU algorithm, the zeroed coefficients remain at zero. When using the EM algorithm, the update rules do not allow the coefficients of $\mathbf{H}$ to be strictly null, hence, we set these coefficients to the eps value in our Matlab® implementation. Observations confirm that these coefficients remain small throughout all the decomposition.

3.4 Summarizing the novelty of the proposed study

While our process is similar in spirit to several existing studies, e.g., [5,8,16], our contribution to the field involves:

- the use of cover multitrack signals instead of hummed or MIDI-synthesis source signals. Our cover signals are expected to provide a more faithful image of the original source signals in the PPM context.

- a stereo NMF framework instead of a mono one. The multichannel framework is expected to exploit spatial information in the demixing process (as far as the convolutive model is a fair approximation of the mixing process). It provides optimal spatial Wiener filters for the separation, as opposed to the {estimated magnitude + mix phase} resynthesis of [8] or the (monochannel) soft masks of [16].

- a synchronization pre-process relying on tempo and musical interpretation instead of, e.g., frame-wise DTW. This is completed with the exploitation of the sources temporal activity for the initialization of $\mathbf{H}$.

4. EXPERIMENTS

4.1 Data and experimental settings

Assessing the performances of source separation on true professionally-produced music data is challenging since the original multitrack signals are necessary to perform objective evaluation but they are seldom available. Therefore, we considered the following data and methodology. The proposed separation algorithm was applied on a series of 4 well-known pop-music songs for which we have the stereo commercial mix signal and two different stereo multitrack covers (see Table 2). The first multitrack cover C1 was provided by iKlax Media, and the second one C2 has been downloaded from the commercial website of another company. We present two testing configurations:

- **Setting 1**: This setting is used to derive objective measures (see below). C1 is considered as the "original multitrack", and used to make a stereo remix of the song which is used as the target mix to be separated. This remix has been processed by a qualified sound engineer with a 10-year background in music

Tracks duration	30 s
Number of channels	I=2
Sampling Rate	32 kHz
STFT frame size	2048
STFT overlap	50 %
Number of iterations	500
Number of NMF components	12 or 50

Table 1: Experimental settings

production, using Avid ProTools®.[3] C2 is considered as the cover version and is used to separate the target mix made with C1.

- **Setting 2**: The original commercial mix is separated using C1 as the cover. This setting is used for subjective evaluation in real-world configuration.

The covers are usually composed of 8 tracks which are quite faithful to the commercial song content as explained in the introduction. For simplicity we merged the tracks to obtain 4 to 6 source signals.[4] All signals are resampled at 32kHz, since source separation above 16kHz has very poor influence on the quality of separated signals and this enables to reduce computations. The experiments are carried out on 30s excerpts of each song.

It is difficult to evaluate the proposed method in reference to existing source separation methods since the cover information is very specific. However, in order to have a reference, we also applied the algorithm with a partial initialization: the spectral patterns $\mathbf{W}$ are here initialized with the cover spectral patterns, whereas the time activation vectors $\mathbf{H}$ are randomly initialized (vs. NMF initialization in the full cover-informed configuration). This enables to i) separate the contribution of cover temporal information, and ii) simulate a configuration where a dictionary of spectral bases is provided by an external database of instruments and voices. This was performed for both EM and MU algorithms. The main technical experimental parameters are summarized in Table 1.

4.2 Separation measures

To assess the separation performances in Setting 1, we computed the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and source image-to-spatial distortion ratio (ISR) defined in [18]. We also calculated the input SIR (SIR_{in}) defined as the ratio between the power of the considered source and the power of all the other sources in the mix to be separated. We consider this criterion because all sources do not contribute to the mix with the same power. Hence, a source with high SIR_{in} is easier to extract than a source with a low SIR_{in}, and SIR_{in} is used to characterize this difficulty.

[3] The source images are here the processed version of C1 just before final summation, hence we do not consider post-summation (non-linear) processing. The consideration of such processing in ISS, as in, e.g., [17], is part of our current efforts.

[4] The gathering was made according to coherent musical sense and panning, e.g., grouping two electric guitars with the same panning in a single track. It is necessary to have the same number of tracks between an original version and its cover. Furthermore, original and cover sources should share approximately the same spatial position (e.g., a cover version of a left panned instrument should not be right panned!)

Title	Tracks	Track names
I Will Survive	6	Bass, Brass, Drums, ElecGuitar, Strings, Vocal.
Pride and Joy	4	Bass, Drums, ElecGuitar, Vocal.
Rocket Man	6	Bass, Choirs, Drums, Others, Piano, Vocal.
Walk this Way	5	Bass, Drums, ElecGuitar1, ElecGuitar2, Vocal.

Table 2: Experimental dataset

Method	SDR	ISR	SIR	SAR
EM W_{init}	0,04	3,51	-1,96	4,82
EM Cover-based	2.45	6.58	4.00	5.38
EM Improvement	2,41	3,08	5,97	0,56
MU W_{init}	-0,98	3,58	-1,14	3,40
MU Cover-based	1.38	6.83	5.04	2.95
MU Improvement	2,36	3,24	6,18	-0,45

Table 3: Average source separation performance for 4 PPM mixtures of 4 to 6 sources (dB).

5. RESULTS

5.1 Objective evaluation

Let us first consider the results obtained with Setting 1. The results averaged across all sources and songs are provided in Table 3. The maximal average separation performance is obtained with the EM cover-informed algorithm with SDR = 2.45dB and SIR = 4.00dB. This corresponds to a source enhancement of $\text{SDR} - \text{SIR}_{\text{in}}$ = 10.05dB and $\text{SIR} - \text{SIR}_{\text{in}}$ = 11.60dB, with the average global SIR_{in} being equal to -7.60dB. These results show that the overall process leads to fairly good source reconstruction and rejection of competing sources. Figure 1a illustrates the separation performances in terms of the difference $\text{SDR} - \text{SIR}_{\text{in}}$ for the song "I will survive". The separation is very satisfying for tracks with sparse temporal activity such as Brass. The Strings track, for which the point source assumption is less relevant, obtains correct results, but tends to spread over other sources images such as Bass. Finally, when cover tracks musically differ from their original sources, the separation performance decreases. This is illustrated with the Electric Guitar (EGtr) and Bass tracks, which do not fully match the original interpretation.

Let us now discuss the cover informed EM and MU methods in relation to the initialization of spectral bases only, referred to as W_{init}. The cover-based EM algorithm provides a notable average SDR improvement of 2.41dB

over EM with W_{init} initialization, and a quite large improvement in terms of SIR (+5.97dB), hence a much better interference rejection. The cover-based MU algorithm also outperforms the MU W_{init} configuration to the same extent (e.g., +2.36dB SDR and +6.18dB SIR improvement). This reveals the ability of the method to exploit not only spectral but also temporal information provided by covers.

Note that both cover-based and W_{init} EM methods outperform the corresponding MU methods in terms of SDR. However, it is difficult to claim for clear-cut EM's better use of the inter-channel mutual information, since EM is slightly lower than MU for SIR (approx. 4dB vs. 5dB for the cover-informed method). In fact, the multichannel framework can take advantage of both spectral and spatial information for source extraction, but this depends on the source properties and mixing configuration. In the song "Walk this way", which detailed results are given in Figure 1b, all sources but the Electric Guitar 1 (Egtr1) are panned at the center of the stereo mixture. Thus, the SDR $-$ SIR_{in} obtained for Egtr1 reaches 20.32dB, as the algorithm relies strongly on spatial information to improve the separation. On the other hand, the estimated Vocal track in "I will survive" is well separated (+8.57dB SDR $-$ SIR_{in} for the cover-informed EM) despite being centered and coincident to other tracks such as Bass, Drums and Electric Guitar (EGtr). In this case, the proposed multichannel NMF framework seems to allow separation of spatially coincident sources with distinct spectral patterns. Depending on the song, some sources obtain better SDR results with the MU algorithm. For example, in "Walk this way", the SDR $-$ SIR_{in} for the Drums track increased from 6.59dB with the EM method to 9.74dB with the MU method. As pointed out in [11], the point source assumption certainly does not hold in this case. The different elements of the drums are distributed between both stereo channels and the source image cannot be modeled efficiently as a convolution of a single point source. By discarding a large part of the inter-channel information, the MU algorithm gives better results in this case. Preliminary tests using a monochannel NMF version of the entire algorithm (monochannel separation using monochannel initialization, as in, e.g., [8, 16]), even show slightly better results for the Drums track, confirming the irrelevancy of the point source convolutive model in this case.

Finally, it can be mentioned that the number of NMF components per source K_j does not influence significantly the SDR and SIR values, although we perceive a slight improvement during subjective evaluation for $K_j = 50$. [5]

5.2 Discussion

Informal listening tests on the excerpts from Setting 2 confirm the previous results and show the potential of cover-informed methods for commercial mix signal separation. [6] Our method gives encouraging results on PPM when point source and convolutive assumptions are respected. For instance, the vocals are in most cases suitably separated, with only long reverberation interferences. As expected, the quality of the mix separation relies on the quality and faithfulness of the cover. A good point is that when original and cover interpretations are well matched, the separated signal sounds closer to the original than to the cover, revealing the ability of the adapted Wiener filters to well preserve the original information.

Comparative experiments with spectral basis initialization only (W_{init}) confirm the importance of the temporal information provided by covers, Although this has not been tested formally, the cover-to-mix alignment of Section 3.2 was shown by informal tests to also contribute to good separation performances.

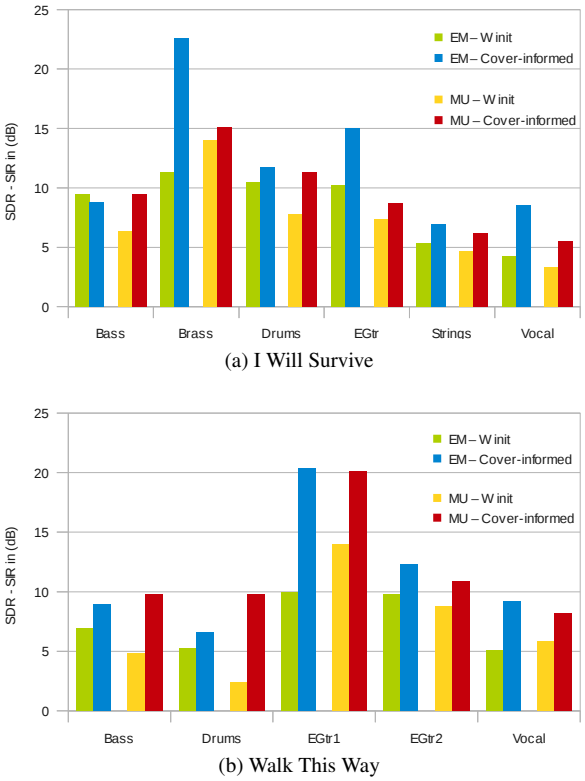

Figure 1: Separation results

6. CONCLUSION

The results obtained by plugging the cover-informed source separation concept in the framework of [11] show that both spectral and temporal information provided by cover signals can be exploited for source separation. This study indicates the interest (and necessity) of using high-quality covers. In this case, the separation process may better take into consideration the music production subtleties, compared to MIDI- or hummed-informed techniques.

Part of the results show the limitations of the convolutive mixing model in the case of PPM. This is the case for sources that cannot be modeled efficiently as a point source convolved on each channel with a linear filter, such as large instruments (e.g., drums and piano). Also, some

[5] Assessing the optimal number of components for each source is a challenging problem left for future work.

[6] Examples of original and separated signals are available at http://www.gipsa-lab.grenoble-inp.fr/∼laurent.girin/demo/ismir2012.html.

tracks such as vocals make use of reverberation times much higher than our analysis frame. As a result, most of the vocals reverberation is not properly separated. The present study and model also do not consider the possible nonlinear processes applied during the mixing process.

Therefore, further research directions include the use of more general models for both sources and spatial processing. For instance, we plan to test the full-rank spatial covariance model of [2], within the very recently proposed general framework of [13] which also enables more specific source modeling, still in the NMF framework (e.g., source-filter models). Within such general model, sources actually composed of several instruments (e.g., drums) may be spectrally and spatially decomposed more efficiently and thus better separated.

7. REFERENCES

[1] S. Dubnov. Optimal filtering of an instrument sound in a mixed recording using harmonic model and score alignment. In *Int. Computer Music Conf. (ICMC)*, Miami, FL, 2004.

[2] N. Q. K. Duong, E. Vincent, and R. Gribonval. Underdetermined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. on Audio, Speech, and Language Proc.*, 18(7):1830–1840, 2010.

[3] J. Engdegård, C. Falch, O. Hellmuth, J. Herre, J. Hilpert, A. Hölzer, J. Koppens, H. Mundt, H. Oh, H. Purnhagen, B. Resch, L. Terentiev, M. Valero, and L. Villemoes. MPEG spatial audio object coding—the ISO/MPEG standard for efficient coding of interactive audio scenes. In *129th Audio Engineering Society Convention*, San Francisco, CA, 2010.

[4] S. Ewert and M. Müller. Score-informed voice separation for piano recordings. In *Proc. of the 12th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Miami, USA, 2011.

[5] S. Ewert and M. Müller. Using score-informed constraints for NMF-based source separation. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Kyoto, Japan, 2012.

[6] C. Faller, A. Favrot, Y-W Jung, and H-O Oh. Enhancing stereo audio with remix capability. In *Proc. of the 129th Audio Engineering Society Convention*, 2010.

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

[8] J. Ganseman, P. Scheunders, G. Mysore, and J. Abel. Source separation by score synthesis. In *Proc. of the Int. Computer Music Conf. (ICMC)*, New-York, 2010.

[9] S. Gorlow and S. Marchand. Informed source separation: Underdetermined source signal recovery from an instantaneous stereo mixture. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2011.

[10] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8):1937–1949, 2012.

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

[12] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, Prague, Czech Republic, 2011.

[13] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Trans. on Audio, Speech and Language Proc.*, 20(4):1118–1133, 2012.

[14] M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Trans. on Audio, Speech, and Language Proc.*, 19(6):1721–1733, 2011.

[15] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Trans. on Audio, Speech, and Language Proc.*, 18(6):1464–1475, 2010.

[16] P. Smaragdis and G. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2009.

[17] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet. Linear mixing models for active listening of music productions in realistic studio condition. In *Proc. of the 132th Audio Engineering Society Conv.*, Budapest, Hungary, 2012.

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

[19] M. Vinyes, J. Bonada, and A. Loscos. Demixing commercial music productions via human-assisted time-frequency masking. In *Proc. of the 120th Audio Engineering Society Convention*, 2006.

[20] J. Woodruff, B. Pardo, and R. B. Dannenberg. Remixing stereo music with score-informed source separation. In *Int. Society for Music Information Retrieval Conference (ISMIR)*, Victoria, Canada, 2006.