Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation

Alexey Ozerov, Cédric Févotte, Raphaël Blouet, Jean-Louis Durrieu

MULTICHANNEL NONNEGATIVE TENSOR FACTORIZATION WITH STRUCTURED CONSTRAINTS FOR USER-GUIDED AUDIO SOURCE SEPARATION

Alexey Ozerov [1], Cédric Févotte [2], Raphaël Blouet [3] and Jean-Louis Durrieu [4]

[1]INRIA, Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes cedex, France
[2]CNRS LTCI; Télécom ParisTech, 75014 Paris, France
[3]Yacast, 4 rue Paul Valery, 75016 Paris, France
[4]Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Separating multiple tracks from professionally produced music recordings (PPMRs) is still a challenging problem. We address this task with a user-guided approach in which the separation system is provided segmental information indicating the time activations of the particular instruments to separate. This information may typically be retrieved from manual annotation. We use a so-called multichannel nonnegative tensor factorization (NTF) model, in which the original sources are observed through a multichannel convolutive mixture and in which the source power spectrograms are jointly modeled by a 3-valence (time/frequency/source) tensor. Our user-guided separation method produced competitive results at the 2010 Signal Separation Evaluation Campaign, with sufficient quality for real-world music editing applications.

Index Terms— Audio source separation, user-guided, nonnegative tensor factorization, generalized expectation maximization.

1. INTRODUCTION

This paper considers multichannel audio source separation from convolutive mixtures, possibly underdetermined (more sources than sensors). We are specifically interested with the challenging problem of professionally produced music recording (PPMRs) separation. This task consists in separating individual tracks (parts played by individual instruments) from multichannel (typically, stereo) music recordings. Providing efficient solutions for the PPMR separation problem can facilitate a wide range of music editing applications including post-production (e.g., stereo to 5.1 upmixing of old recordings) and active-listening.

In general, the PPMR separation problem cannot be solved in a fully blind setting due to the following reasons. First, in PPMRs some sources are often panned in the same direction, and this tendency is observed more and more in modern recordings. Thus, spatial diversity cannot be used alone to separate such sources from each other. Second, some sources do not follow the traditionnal point source assumption. E.g., the drums track is often a sum of several drum elements (bass, snare, hi-hat, etc.) mixed in different directions. Given the such ill-posed nature of the PPMR separation problem, additional information is required to achieve efficiency. As such, some state-of-the-art approaches have addressed less ambitious sub-problems, e.g., extraction of lead singing voice [1]. Other approaches rely on additional information that can be available, e.g., musical score sheet [2], or separated sources [3]. Finally, there are so-called *user-guided* approaches that rely on information manually input by an operator during or prior to the separation process. As such, the source directions are manually selected in [4] and the source to be extracted is hummed in [5].

This paper introduces a novel user-guided approach, where, prior to separation, the user is asked to segment the recording into homogeneous parts containing the target instruments, e.g., vocals/drums, vocals/drums/piano, etc. A similar strategy was considered in [1, 6] for single channel source separation, in the limited cases of either two sources [1] or more than two sources but with a very specific segmentation structure [6]. We extend this strategy to the multichannel case for any given segmentation.

In our framework the segmental information is input to the system through structure constraints in the activation coefficients of the sources model, here a nonnegative tensor factorization (NTF) model. As the sources are not observed directly but only through a multichannel mixture, we term our approach *multichannel NTF* and we aim at estimating both the mixing and source model parameters. This model was proposed in [3] and generalizes our multichannel nonnegative matrix factorization (NMF) model proposed in [7]. Besides the contribution regarding the use of segmental information in the multichannel setting, this paper describes a novel generalized expectation maximization (GEM) algorithm based on multiplicative updates (MU), referred to as *GEM-MU*, that exhibits faster convergence than a GEM algorithm derived from our previous work [7].

The paper is organized as follows. Section 2 describes the source and mixing models. Section 3 describes GEM-MU for maximum likelihood estimation. Section 4 describes how to include the segmental structure into the separation process. Section 5 reports real-world music separation results (with stereo to 5.1 upmix results) and discuss on our results obtained at the 2010 Signal Separation Evaluation Campaign (SiSEC) [8]. Section 6 concludes.

2. MULTICHANNEL NTF MODEL

2.1. Mixing model

We assume that J unknown signals (*the sources*) have been convolutively mixed through I *channels* to produce I observed signals (*the mixtures*). With the standard *narrowband approximation*, this mixing can be expressed in the Short-Time Fourier Transform (STFT) domain as

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

This work was supported in part by the Quaero Programme, funded by OSEO, and by the French ANR projects SARAH (StAndardisation du Remastering Audio Haute-Définition) and TANGERINE (Theory and applications of nonnegative matrix factorization).

where $\mathbf{x}_{fn} = [x_{1fn}, \ldots, x_{Ifn}]^T$ and $\mathbf{s}_{fn} = [s_{1fn}, \ldots, s_{Jfn}]^T$ are the vectors of complex-valued STFTs of the mixtures and the sources, respectively, and $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J}$ is the frequency-dependent mixing matrix ($f = 1, \ldots, F$ is a frequency bin index, $n = 1, \ldots, N$ is a time frame index). $\mathbf{b}_{fn} = [b_{1fn}, \ldots, b_{Ifn}]^T$ represents multichannel residual noise, modeled as zero-mean isotropic Gaussian with covariance $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma_f^2 \mathbf{I}_I$.

2.2. Source model

The 3-valence tensor of source STFTs $\mathbf{S} = \{s_{jfn}\}_{jfn}$, of size $J \times F \times N$, is modeled as a sum of K complex-valued latent tensor components $\mathbf{C}_k = \{c_{k,jfn}\}_{jfn}$. The entries of these components are modeled as the realizations of independent proper complex zero-mean Gaussian random variables with variances $q_{jk} w_{fk} h_{nk}$ ($q_{jk}, w_{fk}, h_{nk} \in \mathbb{R}_+$), such that

$$s_{jfn} = \sum\nolimits_{k=1}^{K} c_{k,jfn}, \qquad c_{k,jfn} \sim \mathcal{N}_c(0, q_{jk} w_{fk} h_{nk}). \quad (2)$$

Thanks to the Gaussian and independence assumptions, the model may also be rewritten as

$$s_{jfn} \sim \mathcal{N}_c(0, v_{jfn}), \qquad v_{jfn} = \sum\nolimits_{k=1}^{K} q_{jk} w_{fk} h_{nk}. \quad (3)$$

Let us introduce the matrices $\mathbf{W} = [w_{fk}]_{fk} \in \mathbb{R}_+^{F \times K}$, $\mathbf{H} = [h_{nk}]_{nk} \in \mathbb{R}_+^{N \times K}$ and $\mathbf{Q} = [q_{jk}]_{jk} \in \mathbb{R}_+^{J \times K}$. Each component $\mathbf{C}_k$ is characterized by the spectral shape given by the k^{th} column of $\mathbf{W}$, with amplitude modulations through frames described by the k^{th} column of $\mathbf{H}$. The columns of $\mathbf{Q}$ model the possible couplings between the components. The columns of $\mathbf{W}$ and rows of $\mathbf{Q}$ are assumed normalized (so that, e.g., they sum to 1), relegating all scale information into $\mathbf{H}$. This general model was proposed in [3] and extends our multichannel NMF model of [7]. Indeed, in [7] we assumed separate NMF model for each of the sources, equivalent to assuming only one nonzero coefficient per row of $\mathbf{Q}$ while this assumption is now relaxed. The interested reader may also refer to [9] for related discussions. Let us also mention that our setting is different from [3] which considers a simpler "informed" source separation application in which the parameters $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ are learnt from the original sources $\mathbf{S}$, assumed available. In our case only the mixtures $\mathbf{X} = \{x_{ifn}\}_{ifn}$ are known.

2.3. Maximum likelihood criterion

Let us denote by $\boldsymbol{\theta}$ the set of model parameters $\{\mathbf{A}, \mathbf{\Sigma}_\mathbf{b}, \mathbf{Q}, \mathbf{W}, \mathbf{H}\}$. The minus log-likelihood function of the parameters writes (up to irrelevant constants)

$$C_\mathbf{X}(\boldsymbol{\theta}) = \sum\nolimits_{fn} \text{trace}\left(\left[\mathbf{x}_{fn}\, \mathbf{x}_{fn}^H \right] \mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \right) + \log\det \mathbf{\Sigma}_{\mathbf{x},fn}, \quad (4)$$

where $\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f (\sum_k \mathbf{\Sigma}_{\mathbf{c},kfn}) \mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f}$ with $\mathbf{\Sigma}_{\mathbf{c},kfn} = \text{diag}([q_{jk}]_j) w_{fk} h_{nk}$.

2.4. Components reconstruction

Given an estimate of $\boldsymbol{\theta}$, the multichannel contribution of the k^{th} latent component to the mixture $\mathbf{c}_{k,fn}^{\text{im}} = \mathbf{A}_f \mathbf{c}_{k,fn}$ (with $\mathbf{c}_{k,fn} = [c_{k,1fn}, \ldots, c_{k,Jfn}]^T$), refered to as *component image*, can be estimated via Wiener filtering such that

$$\hat{\mathbf{c}}_{k,fn}^{\text{im}} = \mathbf{A}_f \mathbf{\Sigma}_{\mathbf{c},kfn} \mathbf{A}_f^H \mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \mathbf{x}_{fn}. \quad (5)$$

The decomposition is conservative, i.e., $\sum_k \hat{\mathbf{c}}_{k,fn}^{\text{im}} + \hat{\mathbf{b}}_{fn} = \mathbf{x}_{fn}$, where $\hat{\mathbf{b}}_{fn} = \mathbf{\Sigma}_{\mathbf{b},fs} \mathbf{\Sigma}_{\mathbf{x},fn}^{-1} \mathbf{x}_{fn}$ is the residual noise estimate.

3. GEM-MU ALGORITHM

We now describe a GEM algorithm for minimization of the likelihood objective function (4). The algorithm is similar in spirit to the one in [7], except that we here consider a reduced latent data set, namely $\mathbf{S}$ (size $J \times F \times N$) instead of $\{\mathbf{C}_k\}_{k=1}^K$ (total size $K \times J \times F \times N$). Note that if the power spectrograms $p_{jfn} = |s_{jfn}|^2$ of the true source were available, as in [3], then maximum likelihood estimation of $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{H}$ would amount to minimize

$$C_\mathbf{S}(\mathbf{Q}, \mathbf{W}, \mathbf{H}) = \sum\nolimits_{jfn} d_{IS}\left(p_{jfn} \left| \sum\nolimits_{k=1}^{K} q_{jk} w_{fk} h_{nk} \right. \right), \quad (6)$$

where $d_{IS}(x|y) = x/y - \log(x/y) - 1$ is the Itakura-Saito divergence, as explained in [9]. Given that the quantities p_{jfn} are not observed directly, they are basically replaced by their posterior value given $\boldsymbol{\theta}$ and data $\mathbf{X}$. As such, one iteration of the GEM algorithm, sketched below, essentially consists of 1) updating the mixing parameters, noise covariance and source power spectrogram posterior estimates $\hat{p}_{jfn}$, 2) update the source model parameters by minimizing (6) with $p_{jfn} = \hat{p}_{jfn}$, which can efficiently be achieved with multiplicative NTF updates [9].

One iteration of GEM-MU :

- **E step**. Conditional expectations of natural statistics:

$$\widehat{\mathbf{R}}_{\mathbf{xx},f} = \frac{1}{N} \sum_n \mathbf{x}_{fn} \mathbf{x}_{fn}^H, \quad \widehat{\mathbf{R}}_{\mathbf{xs},f} = \frac{1}{N} \sum_n \mathbf{x}_{fn} \hat{\mathbf{s}}_{fn}^H, \quad (7)$$

$$\widehat{\mathbf{R}}_{\mathbf{ss},fn} = \hat{\mathbf{s}}_{fn} \hat{\mathbf{s}}_{fn}^H + (\mathbf{I}_J - \mathbf{G}_{\mathbf{s},fn} \mathbf{A}_f) \mathbf{\Sigma}_{\mathbf{s},fn}, \quad (8)$$

$$\widehat{\mathbf{R}}_{\mathbf{ss},f} = \frac{1}{N} \sum_n \widehat{\mathbf{R}}_{\mathbf{ss},fn}, \qquad \hat{p}_{jfn} = \widehat{\mathbf{R}}_{\mathbf{ss},fn}(j,j), \quad (9)$$

  where

$$\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn} \mathbf{x}_{fn}, \qquad \mathbf{G}_{\mathbf{s},fn} = \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H \mathbf{\Sigma}_{\mathbf{x},fn}^{-1}, \quad (10)$$

$$\mathbf{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f}, \quad (11)$$

$$\mathbf{\Sigma}_{\mathbf{s},fn} = \text{diag}\left(\left[\sum\nolimits_{k=1}^{K} q_{jk} w_{fk} h_{nk} \right]_j \right). \quad (12)$$

- **M step**. Update the parameters:

$$\mathbf{A}_f = \widehat{\mathbf{R}}_{\mathbf{xs},f} \widehat{\mathbf{R}}_{\mathbf{ss},f}^{-1}, \quad (13)$$

$$\mathbf{\Sigma}_{\mathbf{b},f} = \text{trace}\left(\widehat{\mathbf{R}}_{\mathbf{xx},f} - \widehat{\mathbf{R}}_{\mathbf{xs},f} \widehat{\mathbf{R}}_{\mathbf{ss},f}^{-1} \widehat{\mathbf{R}}_{\mathbf{xs},f}^H \right) \mathbf{I}_I / I, \quad (14)$$

$$\mathbf{Q} = \mathbf{Q} \odot \frac{< \mathbf{V}^{\odot -2} \odot \widehat{\mathbf{P}}, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}}}{< \mathbf{V}^{\odot -1}, \mathbf{W} \circ \mathbf{H} >_{\{2,3\},\{1,2\}}}, \quad (15)$$

$$\mathbf{W} = \mathbf{W} \odot \frac{< \mathbf{V}^{\odot -2} \odot \widehat{\mathbf{P}}, \mathbf{Q} \circ \mathbf{H} >_{\{1,3\},\{1,2\}}}{< \mathbf{V}^{\odot -1}, \mathbf{Q} \circ \mathbf{H} >_{\{1,3\},\{1,2\}}}, \quad (16)$$

$$\mathbf{H} = \mathbf{H} \odot \frac{< \mathbf{V}^{\odot -2} \odot \widehat{\mathbf{P}}, \mathbf{Q} \circ \mathbf{W} >_{\{1,2\},\{1,2\}}}{< \mathbf{V}^{\odot -1}, \mathbf{Q} \circ \mathbf{W} >_{\{1,2\},\{1,2\}}}, \quad (17)$$

  where $\mathbf{V} = \{\hat{v}_{jfn}\}_{jfn}$, $\widehat{\mathbf{P}} = \{\hat{p}_{jfn}\}_{jfn}$, $\odot$ denotes element-wise operation (the division $\cdot/\cdot$ is here element-wise as well), $\mathbf{A} \circ \mathbf{B}$ is the $F \times N \times K$ tensor with elements $a_{fk} b_{nk}$ (similar to Khatri-Rao product), and $< \mathbf{S}, \mathbf{T} >_{\mathcal{K}_S, \mathcal{K}_T}$ denotes the contracted product between tensors $\mathbf{S}$ and $\mathbf{T}$, defined in Appendix B of [9].

- Normalize $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$, following section 2.2 or [7].

4. USER-GUIDED SEPARATION

So far we have kept the level of exposition of the source separation system at a general level and have not made specific structure assumptions on any of the model parameters. Structure constraints can easily be set on $\mathbf{Q}$, $\mathbf{W}$ or $\mathbf{H}$ in the forms of zero coefficients, because these coefficients will be unchanged under the multiplicative updates. The proposed method describes as follows.

1. The user (manually) performs a temporal segmentation of the tracks to separate and decides on the number of components per track (e.g., 2 for bass, 7 for vocals, etc.)
2. The temporal segmentation and the number of components per track are reflected in $\mathbf{H}$ in the forms of zeros. E.g., if source 1 is assigned two first components and is silent between frames 100 and 200, then we set $h_{n1} = h_{n2} = 0$ for $n = 100 \ldots 200$. The other coefficients are randomly initialized to positive values.
3. The remaining parameters are initialized using and ad-hoc procedure described in Section IV.H of [7], based on NMF of the stacked channel spectrograms (initialization of $\mathbf{W}$) and clustering of the spatial cues (initialization of $\mathbf{A}$ and $\mathbf{Q}$).[1]
4. Run the GEM-MU algorithm of Section 3.
5. Compute estimates of every component $\hat{\mathbf{c}}_{k,fn}^{\mathrm{im}}$ with Eq. (5).
6. The operating sound engineer may listen to reconstructed components, and if some sources share the exact same temporal segmentation (e.g., when two sources always play together) the operator may manually fine-tune the components grouping.

5. EXPERIMENTS

5.1. Convergence of GEM-MU vs. GEM

We have applied the proposed GEM-MU algorithm and the GEM algorithm from [7] to 4 linear instantaneous mixtures and 8 synthetic convolutive mixtures of 3 sources from the SiSEC 2010 "Underdetermined speech and music mixtures" task development dataset. Figure 1 shows the resulting average cost functions (4). GEM-MU converges faster than GEM with respect to the likelihood value. The computational complexities of one iteration of GEM-MU and GEM are comparable.

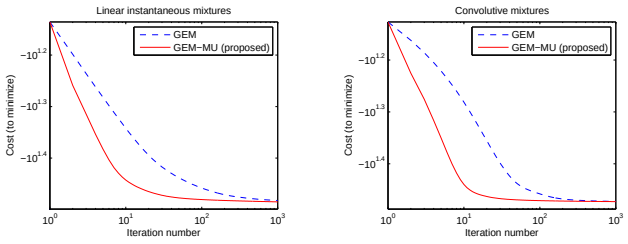

Fig. 1. Comparison between GEM-MU (proposed) and GEM in terms of average cost value for linear instantaneous (left) and synthetic convolutive (right) mixtures, as a function of the iteration number.

[1] As a matter of fact the initialization provides a sparse matrix $\mathbf{Q}$ which fixes the repartition of components per source, amounting to our initial multichannel NMF model [7]. Full update the label matrix $\mathbf{Q}$ under sparse constraints is left for future work. See also discussion in [9].

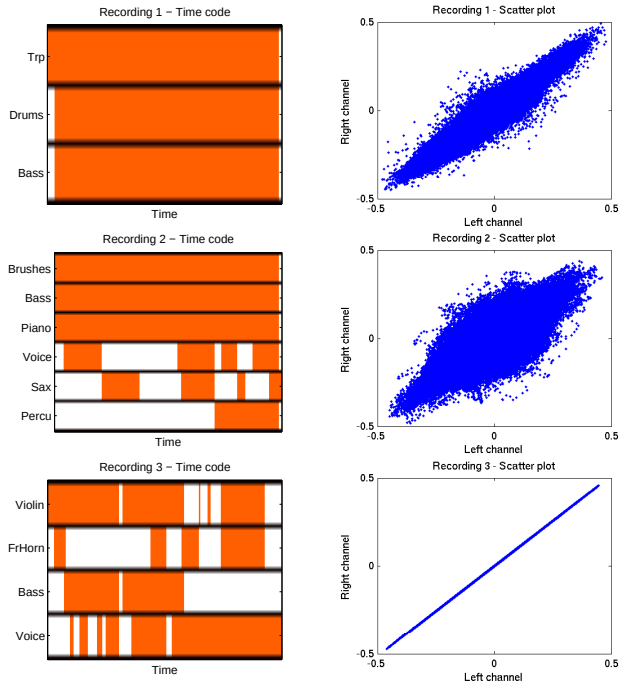

Fig. 2. Time-codes (orange = track is active) (left) and scatter plots (left vs. right channel samples) (right) of three PPMRs.

5.2. Separation and upmixing to 5.1

Our method was applied to three recordings having the following different characteristics: the first one has a poor segmental diversity (much overlap between the sources in time), but rich spatial diversity, the second one has both rich segmental and spatial diversity, and the third one has rich segmental diversity, but poor spatial diversity (it is a mono recording), see Fig. 2. The estimated tracks were then remixed to 5.1 by professional sound engineers in the context of the ANR-SARAH project.[2] Objective source separation evaluation could not be performed as the original sources are not available for these recordings. As such, we set a companion webpage were audio samples can be listened to, including results with a user-guided version of the DUET algorithm [11] for comparison (stereo data only).[3] Informal listening shows that, as compared to results by the DUET algorithm, separation results obtained by our method do not suffer from disturbing "musical noise" artefacts. Moreover, DUET was not able to separate the tracks of the 3rd (mono) recording, as well as some tracks from the 1st and 2nd recordings, possibly because they are mixed in the same direction. Note that the manual intervention required by our approach is moderate, typically not more than 40 minutes are needed to annotate a 5 minute-long recording, and this could certainly be reduced with a dedicated annotation software.

5.3. SiSEC 2010 results

Finally, we report results of the "PPMR" separation task of SiSEC 2010 [8], to which our method of Section 4 entered. Only three methods entered the task and all involved some degree of user-guidance. The results are displayed Table 1. Our results are com-

[2] http://sarah.audionamix.com/

[3] http://www.irisa.fr/metiss/ozerov/multi_ntf_demo.html

		Glen Philips "The Spirit of Shackleton"			Nine Inch Nails "The Good Soldier"		Shannon Hurley "Sunrise"				Average
		vocals	drums	bass	vocals	drums	vocals	drums	bass	piano	
Algorithm 1	SDR (dB)	3.3	2.3	-4.0	1.1	5.7	2.2	3.6	2.6	-2.3	1.6
Proposed	OPS (0-100)	19.5	31.9	14.3	30.2	27.9	15.5	39.2	8.3	18.1	22.7
Algorithm 2	SDR (dB)	-0.3	-	-	-2.6	-	0.8	-	-	-	-
J. Janer & R. Marxer [4]	OPS (0-100)	15.9	-	-	18.8	-	15.2	-	-	-	-
Algorithm 3	SDR (dB)	3.9	3.6	-2.0	1.1	1.2	2.2	4.7	3.4	-3.8	1.6
M. Spiertz [10]	OPS (0-100)	15.4	37.3	8.7	25.2	25.0	8.0	40.6	5.8	10.4	19.6
STFT	SDR (dB)	5.6	6.4	2.0	1.5	5.1	7.8	7.3	8.2	0.7	4.9
Ideal Binary Mask	OPS (0-100)	21.0	30.6	11.3	29.3	37.9	15.4	35.5	20.6	19.2	24.5
Cochleagram	SDR (dB)	3.9	1.1	0.7	1.4	1.6	6.1	1.5	1.6	0.4	2.0
Ideal Binary Mask	OPS (0-100)	15.7	26.2	15.6	17.6	37.5	11.2	42.8	30.4	12.1	23.2

Table 1. SiSEC 2010 PPMRs task results (test2 dataset with full-length recordings).

parable to Spiertz's in terms of average Source to Distoration Ratio (see [12] for description) and slightly better in terms of average overall perceptual score (OPS), a newly proposed auditory motivated measure [13]. Interestingly, ideal binary masks (using the ground truth) only give a small improvement in terms of average OPS as compared to our method, illustrating the validity of our model together with Wiener filter-based estimates.

6. CONCLUSION

We have presented a novel user-guided audio source separation method based on a multichannel NTF source model that can easily include structure constraints. Also, the approach is general enough to accomodate various mixing hypotheses (underdetermined/overdetermined, instantaneous/narrowband convolutive). In particular, our method is suited to the challenging PPMR separation task, which has been the primary objective of this paper. The ill-posed nature of PPMR separation imposes some level of user-guidance, consisting in our case of manual annotations. This is a reasonable requirement for the music-editing application (mono/stereo to 5.1 upmix) that we considered in this paper.

As for further research, one direction would be to replace user-guided segmentation by an automatic source identification system, or more generally to go towards joint source identification and separation approaches. Another interesting research direction would be, instead of fixing *a priori* structured constraints on multichannel NTF model parameters (i.e., $\mathbf{Q}$ and $\mathbf{H}$), to learn this structure *a posteriori* using some sparsity-inducing constraints.

7. REFERENCES

[1] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[2] J. Ganseman, G. J. Mysore, J.S. Abel, and P. Scheunders, "Source separation by score synthesis," in *Proceedings of the International Computer Music Conference (ICMC)*, New York, NY, June 2010.

[3] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, St Malo, France, 2010.

[4] M. Vinyes, J. Bonada, and A. Loscos, "Demixing commercial music productions via human-assisted time-frequency masking," in *Proceedings of Audio Engineering Society 120th Convention*, 2006.

[5] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proceedings IEEE Workshop Applications of Signal Processing to Audio and Acoustics WASPAA '09*, 2009, pp. 69–72.

[6] H. Laurberg, M. N. Schmidt, M. G. Christensen, and S. H. Jensen, "Structured non-negative matrix factorization with sparsity patterns," in *Proceedings Asilomar Conference on Signals, Systems, and Computers*, 2008.

[7] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Lang. Proc.*, vol. 18, no. 3, pp. 550–563, March 2010.

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

[9] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues," in *7th International Symposium on Computer Music Modeling and Retrieval (CMMR 2010)*, 2010.

[10] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proceedings of International Conference on Digital Audio Effects DAFx'09*, Como, Italy, Sept. 2009.

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

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

[13] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, submitted.