An interactive audio source separation framework based on non-negative matrix factorization

Ngoc Duong, Alexey Ozerov, Louis Chevallier, Joel Sirot

AN INTERACTIVE AUDIO SOURCE SEPARATION FRAMEWORK BASED ON NON-NEGATIVE MATRIX FACTORIZATION

Ngoc Q. K. Duong, Alexey Ozerov, Louis Chevallier, and Joël Sirot

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{quang-khanh-ngoc.duong, firstname.lastname}@technicolor.com

ABSTRACT

Though audio source separation offers a wide range of applications in audio enhancement and post-production, its performance has yet to reach the satisfactory especially for single-channel mixtures with limited training data. In this paper we present a novel interactive source separation framework that allows end-users to provide feedback at each separation step so as to gradually improve the result. For this purpose, a prototype graphical user interface (GUI) is developed to help users annotating time-frequency regions where a source can be labeled as either *active*, *inactive*, or *well-separated* within the displayed spectrogram. This user feedback information, which is partially new with respect to the state-of-the-art annotations, is then taken into account in a proposed uncertainty-based learning algorithm to constraint the source estimates in next separation step. The considered framework is based on non-negative matrix factorization and is shown to be effective even without using any isolated training data.

Index Terms— Interactive audio source separation, non-negative matrix factorization, uncertainty-based learning, time-frequency annotation, user feedback.

1. INTRODUCTION

Audio source separation still remains a very hot research topic even there has been a surge of research over the past years [1]. In fact, separation performance that can be achieved by state-of-the-art algorithms is far from satisfactory in certain scenarios such as under-determined mixtures of reverberated sources. In single-channel case, where spatial information about the sources cannot be exploited, the problem becomes even harder and usually relevant training data is needed to first learn the spectral characteristics of individual sources. Such a class of supervised algorithms is mostly based on non-negative matrix factorization (NMF) [2, 3], or its probabilistic formulation known as probabilistic latent component analysis (PLCA) [4, 5]. However, when training examples are unavailable (or not representative enough) these methods can not be applied without other prior information about the sources. Examples of such prior information include "hummed" sounds that mimic the ones in the mixture [6], or text transcriptions of the corresponding speech sources [7].

So called *user-guided* approaches based on NMF have been proposed recently. These approaches allow end-user to manually annotate information about activity of each sound source, *e.g.* if it is active or not, in time [8] or time-frequency (T-F) [9] domains. The annotated source activity information is then used, instead of training data, to guide the separation process. Another user-guided approach based on independent vector analysis, which allows user to tune temporal power variations of sources, was also presented in [10]. Though prior work has shown the effectiveness of such algorithms, the results are still far from perfect especially in mixing scenarios with strong overlapping sources. A more advanced *interactive* framework based on PLCA was proposed by Bryan *et al.* [11] whereby overall separation process comprises several interactive separation steps (if needed). At each step, end-user can perform T-F annotation on the spectrogram displays of intermediate separation results, in addition to the annotation on the spectrogram of the mixture itself. Note that in contrast to [8, 9], where the annotations are specified only once, interactive approach in [11] allows user feedback at each separation step so as to gradually improve the result by correcting remaining errors.

Motivated by the effectiveness of both the user-guided algorithms [8, 9] and the interactive strategy [11][1], we present in this paper an *interactive source separation approach* based on a novel NMF formulation. The proposed approach can efficiently handle user feedback information at each separation step, and is robust to errors in the annotations thank to the derived uncertainty-based learning algorithm for parameter estimates. In addition to the T-F annotation about the source activity considered in the existing algorithms, we introduce a new type of annotation about the quality of the separated sources: user can validate if a source is *well-separated* in certain T-F regions. These well-separated regions can then be effectively exploited, as sort of training data, in our formulation to better constrain parameter estimates in next separation step.

The rest of the paper is organized as follows. In Section 2 we summarize the NMF model for source separation and a baseline parameter estimate exploiting temporal annotation as in [8]. This baseline can serve as the first separation step in our interactive framework. Main contributions of the paper are described in Section 3 where the T-F annotations via a GUI are introduced followed by the proposed optimization algorithm and a global description of the interactive system. We conduct experiments to validate the effectiveness of the proposed approach in Section 4. Finally we conclude in Section 5.

2. NMF MODEL AND BASELINE PARAMETER UPDATE

2.1. Model

Let x_{fn} and $s_{j,fn}$ be the short-time Fourier transform (STFT) coefficients of the observed single-channel mixture signal and the con-

[1]This approach obtained better objective separation result compared to other submissions in "Professionally produced music recordings" task of the Signal Separation Evaluation Campaign (SiSEC2013): `http://www.onn.nii.ac.jp/sisec13/evaluation_result/MUS/testMUS2013.htm`

tribution of j-th source signal, respectively, at frequency bin f and time frame n. The mixing model writes

$$x_{fn} = \sum_{j=1}^{J} s_{j,fn}, \tag{1}$$

where J is the total number of sources, $f = 1, \ldots, F$ and $n = 1, \ldots, N$. Defining the power spectrogram of the mixture by $v_{fn} = |x_{fn}|^2$, NMF aims at approximately factorizing the $F \times N$ matrix $\mathbf{V} = \{v_{fn}\}_{fn}$ as $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}$, where $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices of size $(F \times K)$ and $(K \times N)$, respectively. Assuming that the source STFT coefficients follow zero-mean Gaussian distribution $s_{j,fn} \sim \mathcal{N}_c(0, \hat{v}_{j,fn})$ where $\hat{v}_{j,fn} = [\mathbf{W}_{(j)}\mathbf{H}_{(j)}]_{fn}$, $\mathbf{W}_{(j)}$ and $\mathbf{H}_{(j)}$ are matrices of size $(F \times K_j)$ and $(K_j \times N)$, respectively, modeling the contribution of the j-th source. With linear model $\hat{v}_{fn} = \sum_j \hat{v}_{j,fn}$ [3], we can write $\mathbf{W} = [\mathbf{W}_{(1)}, \ldots, \mathbf{W}_{(J)}]$ and $\mathbf{H} = [\mathbf{H}_{(1)}^T, \ldots, \mathbf{H}_{(J)}^T]^T$.

2.2. Baseline parameter update exploiting temporal annotation

NMF parameters can be estimated in the maximum likelihood (ML) sense, which is equivalent to minimizing the following cost function [3]

$$C_1(\boldsymbol{\theta}) = \sum_{fn} d_{IS}\left(v_{fn}|\hat{v}_{fn}\right), \tag{2}$$

where $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ being the NMF parameters, and $d_{IS}(\cdot|\cdot)$ denotes Itakura-Saito (IS) divergence. The well-known multiplicative update (MU) rules for parameter estimation write:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}\right)}{\mathbf{W}^T(\mathbf{WH})^{\cdot -1}} \tag{3}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}\right)\mathbf{H}^T}{(\mathbf{WH})^{\cdot -1}\mathbf{H}^T} \tag{4}$$

where $\odot$ denotes the Hadamard entrywise product, $\mathbf{A}^{\cdot p}$ being the matrix with entries $[\mathbf{A}]_{ij}^p$, and the division is entrywise.

Considering the baseline user-guided approach [8], parameter estimation is guided by temporal annotation where user is asked to indicate time segments along the mixture where each source is active. This temporal annotation is then reflected in the initialization of the activation matrix $\mathbf{H}$: when j-th source is not annotated by "active" at time frame n, the n-th column of $\mathbf{H}_{(j)}$ is initialized by 0 and it will not be changed via the multiplicative update. Note that this temporal annotation is much easier and faster than the detail time-frequency annotation considered in [9, 11] as well as later part of the paper. Thus, with the absence of isolated training sounds, this baseline can serve as the first separation step in our interactive framework.

2.3. Source reconstruction

Given the estimated parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, the source STFT coefficients are computed by Wiener filtering as

$$\hat{s}_{j,fn} = \frac{\hat{v}_{j,fn}}{\hat{v}_{fn}} x_{fn}, \tag{5}$$

then the time domain source estimates are obtained via the inverse STFT.

3. PROPOSED INTERACTIVE SEPARATION ALGORITHM

We first present all types of the considered time-frequency annotations, which would bring benefit to the separation process, in Section 3.1. We then propose an uncertainty-based learning algorithm that allows to efficiently incorporate all annotated information to constrain parameter estimates in interactive separation steps in Section 3.2.

3.1. Time-frequency annotations

In order to support the annotation task, we developed a prototype user interface enabling user to select T-F regions (either in rectangle or polygon shapes) in the spectrogram display (either in linear or logarithmic frequency scale) of the mixture as well as the separated sources. This GUI is visualized in Fig. 1. Given a selected T-F region, and possibly by hearing the corresponding sound, end-user can assign to each source one of the following labels:

(i) source is *active* (default label),

(ii) source is *inactive*,

(iii) source is *well-separated*.

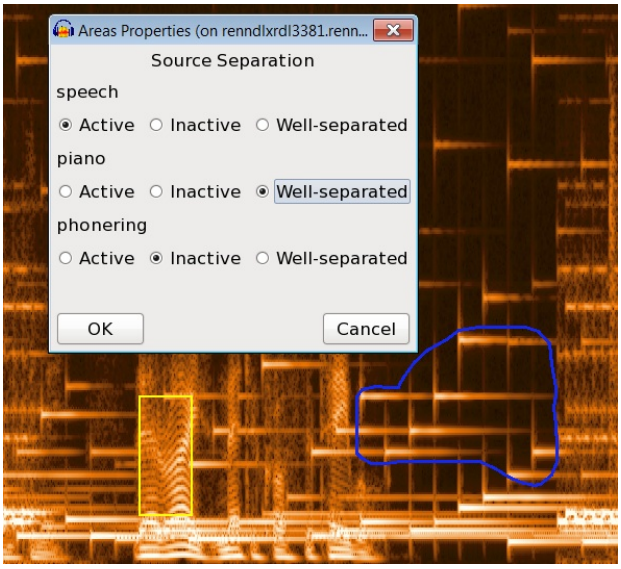

Fig. 1. GUI for user annotation. Piano sound is labeled as *well-separated* in the current selected polygon region (in blue) where phone-ring sound labeled as is *inactive*. User also annotated that speech is *active* alone (*i.e.* all the other sources are inactive) in the rectangle region (in yellow).

While annotation (i) and (ii) have been addressed in [9, 11], annotation (iii) is new. We find that this type of information is interesting because first, the well-separated T-F regions should be constrained so as they will not be accidently damaged due to the changing of parameters in next separation steps. Second, and even more important, these regions can be used, as sort of input training data, to help better estimating the model parameters of the corresponding source.

3.2. Interactive parameter update exploiting T-F annotations

We derive an optimization algorithm addressing two challenges: (1) how to use annotated T-F regions to learn NMF models, and (2) how to handle errors in the annotation. While the former can be solved in the same way as in [9], the latter is more difficult. Also, the latter is crucial so as to account for the *well-separated* regions where artifacts and interferences still remain from the previous separation step[2]. To fulfil this challenge, we exploit uncertainty-based learning [12, 13] where source models can be learned from the source estimates, while taking into account separation errors described by their variances. The principle is summarized as follow. Within the above-mentioned Gaussian assumption, the posterior of j-th source writes [12]

$$p(s_{j,fn}|x_{fn};\boldsymbol{\theta}) = \mathcal{N}_c\left(s_{j,fn}; \hat{s}_{j,fn}, \frac{\hat{v}_{j,fn}\hat{v}_{\bar{j},fn}}{\hat{v}_{fn}}\right), \quad (6)$$

where $\hat{v}_{\bar{j},fn} = \hat{v}_{fn} - \hat{v}_{j,fn}$ and $\hat{s}_{j,fn}$ is the reconstructed source from Wiener filtering (5), which can be then written as

$$\hat{s}_{j,fn} = s_{j,fn} + b_{j,fn}, \quad (7)$$

where $b_{j,fn} \sim \mathcal{N}_c(0, \hat{v}_{j,fn}\hat{v}_{\bar{j},fn}/\hat{v}_{fn})$ is source estimation error. Estimating the NMF model for j-th source from (7) in the ML sense can be shown to be equivalent to minimizing the following cost [12]

$$C_2(\boldsymbol{\theta}) = \sum_{j,fn} d_{IS}\left(\tilde{v}_{j,fn}|\hat{v}_{j,fn} + \tilde{e}_{j,fn}\right), \quad (8)$$

where $\tilde{v}_{j,fn} = (\hat{v}_{j,fn}/\hat{v}_{fn})^2 v_{fn}$ and $\tilde{e}_{j,fn} = \hat{v}_{j,fn}\hat{v}_{\bar{j},fn}/\hat{v}_{fn}$ are computed from model parameters estimated at the previous separation step and fixed at current estimation step. It can be seen that this parameter estimation takes into account both posterior mean $\hat{s}_{j,fn}$ and posterior variance $\tilde{e}_{j,fn}$ characterizing source reconstruction error.

The cost (8) is now used to account for the annotated T-F regions in our framework. Combining (2) and (8), the overall proposed cost function to be optimized for interactive parameter estimates is

$$C_3(\boldsymbol{\theta}) = \sum_{fn} d_{IS}\left(v_{fn}|\hat{v}_{fn}\right) + \sum_{j,fn} \lambda_{j,fn} d_{IS}\left(\tilde{v}_{j,fn}|\hat{v}_{j,fn} + \tilde{e}_{j,fn}\right), \quad (9)$$

where $\lambda_{j,fn}$, $\tilde{v}_{j,fn}$ and $\tilde{e}_{j,fn}$ are defined, for each source j at each time-frequency point (f, n), according to the T-F annotations as:

$$(\lambda_{j,fn}, \tilde{v}_{j,fn}, \tilde{e}_{j,fn}) = \begin{cases} (a, 0, 0) & \text{if } j \text{ is } \textit{inactive}, \\ (b, v_{fn}, 0) & \text{if } j \text{ is } \textit{active alone}, \\ \left(c, \left(\frac{\hat{v}_{j,fn}}{\hat{v}_{fn}}\right)^2 v_{fn}, \frac{\hat{v}_{j,fn}\hat{v}_{\bar{j},fn}}{\hat{v}_{fn}}\right) & \text{if } j \text{ is } \textit{well-separated}, \\ (0, \texttt{any}, \texttt{any}) & \text{otherwise.} \end{cases} \quad (10)$$

where positive constants a, b, and c determine contributions of different annotation types, `any` means any value, and a source is *active alone* when all other sources are labeled as inactive. Note that all parameters $\hat{v}_{j,fn}$, $\hat{v}_{fn}$, and $\hat{v}_{\bar{j},fn}$ in (10) are computed from the previous separation step and fixed in current interactive step. Further explanation for the quantities in (10) is as follow. When a source is labeled as *inactive* or *active alone*, its corresponding separation error variances $\tilde{e}_{j,fn}$ are set to zero, while $\tilde{v}_{j,fn}$ should be equal to either zero or the mixture power spectrum v_{fn}, respectively. Besides, if a source is already *well-separated* in certain regions, its corresponding STFT coefficients should be constrained to keep values from the previous separation step and not to vary greatly.

Let $\boldsymbol{\Lambda}_{(j)}$, $\widetilde{\mathbf{V}}_{(j)}$ and $\widetilde{\mathbf{E}}_{(j)}$ the $F \times N$ matrices characterizing the T-F annotations for j-th source with, respectively, $\lambda_{j,fn}$, $\tilde{v}_{j,fn}$ and $\tilde{e}_{j,fn}$ as entries. Using general principle from [3] we derived MU rules to minimize (9) and the resulting parameter updates are as follows:

$$\mathbf{H}_{(j)} \leftarrow \mathbf{H}_{(j)} \odot \frac{\mathbf{W}_{(j)}^T\Big(\left(\mathbf{W}\mathbf{H}\right)^{\cdot -2} \odot \mathbf{V} + \boldsymbol{\Lambda}_{(j)} \odot \left(\mathbf{W}_{(j)}\mathbf{H}_{(j)} + \widetilde{\mathbf{E}}_{(j)}\right)^{\cdot -2} \odot \widetilde{\mathbf{V}}_{(j)}\Big)}{\mathbf{W}_{(j)}^T\Big(\left(\mathbf{W}\mathbf{H}\right)^{\cdot -1} + \boldsymbol{\Lambda}_{(j)} \odot \left(\mathbf{W}_{(j)}\mathbf{H}_{(j)} + \widetilde{\mathbf{E}}_{(j)}\right)^{\cdot -1}\Big)} \quad (11)$$

$$\mathbf{W}_{(j)} \leftarrow \mathbf{W}_{(j)} \odot \frac{\Big(\left(\mathbf{W}\mathbf{H}\right)^{\cdot -2} \odot \mathbf{V} + \boldsymbol{\Lambda}_{(j)} \odot \left(\mathbf{W}_{(j)}\mathbf{H}_{(j)} + \widetilde{\mathbf{E}}_{(j)}\right)^{\cdot -2} \odot \widetilde{\mathbf{V}}_{(j)}\Big)\mathbf{H}_{(j)}^T}{\Big(\left(\mathbf{W}\mathbf{H}\right)^{\cdot -1} + \boldsymbol{\Lambda}_{(j)} \odot \left(\mathbf{W}_{(j)}\mathbf{H}_{(j)} + \widetilde{\mathbf{E}}_{(j)}\right)^{\cdot -1}\Big)\mathbf{H}_{(j)}^T} \quad (12)$$

3.3. Overall system description

Following the baseline separation step summarized in Section 2.2 and the proposed interactive refinement steps with user's T-F annotations introduced in Section 3.2, this section summarizes the global workflow of the overall system. For clarity, all steps in the proposed interactive framework are presented in Algorithm 1.

Note that all T-F annotations in interactive separation steps are kept in the memory and will be reused, together with new annotations, in next separation step.

4. EXPERIMENTS

We first evaluated the source separation performance of the proposed approach over four single-channel mixtures of different male and female speeches (extracted from TIMIT database) with different background sounds (piano chords, drums, and phone ring). The mixtures were 23 second long, sampled at 16 kHz, and articially mixed at 0 dB SNR. In each mixture speeches (either male or female voices) and background sounds appear alone during about three seconds so that temporal annotation was well-exploited to guide the parameter estimates in the first separation step.

We compared separation result of the derived interactive approach after the first separation step (named Int-SS-1) with that obtained after the second separation step when the proposed annotation about the quality of separated sources, *i.e. well-separated* T-F regions, is either not used (named Int-SS-2) or used (named Int-SS-3). The time-frequency annotation was performed by one of the author who has experience on sound processing. Note that Int-SS-1 is equivalent to the user-guided algorithm [8] where only temporal annotation was used, and Int-SS-2 is comparable to the state-

[2]Though well-separated regions are validated by end-user via listening, they are usually not error-free. Indeed, artifacts and interferences still exist in these regions, but they are simply masked by the target source.

Algorithm 1 Global workflow of the proposed interactive source separation algorithm.

1. User listens to the mixture and roughly annotate temporal segments where each source is active
2. Randomly initialize NMF parameters $\{\mathbf{W}_{(j)}, \mathbf{H}_{(j)}\}_j$, then set n-th column of $\mathbf{H}_{(j)}$ to 0 if j-th source is inactive at time frame n (see Section 2.2)
3. Perform first source separation step: parameters are alternatingly updated by (3) and (4) until convergence, then sources are reconstructed by Wiener filtering (see Section 2.3)
4. User evaluates the separation performance. Stop separation process when the result is satisfied, otherwise:
5. User performs the detailed T-F annotations (see Section 3.1)
6. Construct the $F \times N$ nonnegative matrices $\mathbf{\Lambda}_{(j)}$, $\widetilde{\mathbf{V}}_{(j)}$ and $\widetilde{\mathbf{E}}_{(j)}$ for all source j from all the T-F annotations as (10)
7. Initialize NMF parameters similarly to step 2
8. Perform interactive source separation step: parameters are alternatingly updated by (11) and (12) until convergence, then sources are reconstructed by Wiener filtering
9. Go to step 4.

of-the-art interactive method [11] though the former is formulated based on NMF while the later uses PLCA. For parameter estimation, the number of MU iterations was set to 100 for all separation steps, $K_j = 20, j = 1, 2$, and the NMF parameters $\mathbf{W}_{(j)}, \mathbf{H}_{(j)}$ were re-initialized, in the second separation step, in the same way as they were initialized in the first step. We observed that this re-initialization brings slightly better result than initializing directly by the parameters obtained after MU iterations in the first separation step. The trade-of parameters a, b, and c are set manually between 1 and 10.

Approach	SDR	SIR	SAR
Int-SS-1 (comparable to [8])	8.9	16.0	13.1
Int-SS-2 (comparable to [11])	9.8	16.8	13.1
Int-SS-3 (proposed)	10.2	17.4	13.3

Table 1. Average source separation performance.

Separation performance was evaluated using the signal-to-distortion ratio (SDR) criterion measuring overall distortion, as well as the signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) criteria [14], measured in dB, averaged over all sources, and shown in Table 1. As expected, performing source separation interactively even with only one user-feedback step (Int-SS-2 and Int-SS-3) significantly improves the result, *i.e.* in terms of SDR and SIR, compared to the baseline (Int-SS-1). This is because some interferences remained in separated sources from the first separation step, which were annotated by user, were successfully removed in the second step. Besides, it is worth noticing that Int-SS-3 improves the SDR, SIR, and SAR by 0.4 dB, 0.6 dB, and 0.2 dB, respectively, compared to Int-SS-2. This result confirms the effectiveness of our proposed *well-separated* annotation. In addition to the above synthetic mixtures, we have also tested the proposed approach on a twelve minute real-world sound track of Beverly Hills drama series for the separation of dialogs (speeches from main actors and actresses) and background sounds (music, environmental noise, footstep, etc). Due to the lack of groundtruth signals for objective evaluation, informer listening has confirmed the quality of separated sources after two interactive steps with about ninety minutes of manual annotation.

5. CONCLUSION

In this paper, we have presented a novel interactive source separation framework based on NMF formulation. The proposed approach efficiently exploits a new type of user annotation about the quality of the estimated sources, in addition to the existing T-F annotations about the contributions of the sources to the mixture spectrogram, to further constrain the parameter estimates in interactive separation steps. Preliminary experiments with both simulated and real-world mixtures confirm the effectiveness of the derived algorithm. Future research would be devoted to learn an automatic or semi-automatic annotation algorithm so as to reduce manual effort and fasten annotation time.

6. REFERENCES

[1] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

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

[4] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proc. IEEE Workshop on Aplications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 17 – 20.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 414 – 421.

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

[7] L. L. Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *Proc. Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2013.

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

[9] A. Lefèvre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for singlechannel source separation," in *the Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2012.

[10] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 2417–2420.

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

[12] S. Arberet, A. Ozerov, F. Bimbot, and R. Gribonval, "A tractable framework for estimating and combining spectral source models for audio source separation," *Signal Processing*, vol. 92, no. 8, pp. 1886–1901, 2012.

[13] A. Ozerov, M. Lagrange, and E. Vincent, "Uncertainty-based learning of acoustic models from noisy data," *Computer Speech and Language*, vol. 27, no. 3, pp. 874–894, 2013.

[14] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 734–741.