Temporal annotation-based audio source separation using weighted nonnegative matrix factorization

Ngoc Q. K. Duong, Alexey Ozerov, Louis Chevallier

Temporal annotation-based audio source separation using weighted nonnegative matrix factorization

Ngoc Q. K. Duong, Alexey Ozerov, and Louis Chevallier
Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
Email: {alexey.ozerov, quang-khanh-ngoc.duong, louis.chevallier}@technicolor.com

***Abstract*—We consider an emerging user-guided audio source separation approach based on the temporal annotation of the source activity along the mixture. In this baseline algorithm nonnegative matrix factorization (NMF) is usually used as spectral model for audio sources. In this paper we propose two weighting strategies incorporated in the NMF formulation so as to better exploit the annotation. We then derive the corresponding multiplicative update (MU) rules for the parameter estimation. The proposed approach was objectively evaluated within the fourth community-based Signal Separation Evaluation Campaign (SiSEC 2013) and shown to outperform the baseline algorithm, while obtaining comparable result to some other state-of-the-art methods.**

***Index Terms*—User-guided audio source separation, temporal annotation, nonnegative matrix factorization (NMF), weighted NMF.**

I. Introduction

Audio source separation, which aims at extracting individual sound sources from their observed mixture, offers a wide range of applications in audio enhancement for communication, robotics, and audio post-production. Since blind source separation has remained challenging in many real-world mixtures [1], especially for single-channel mixtures, supervised approaches, which exploit relevant training data to first learn the spectral characteristics of individual sources [2], [3], has been considered instead. However, when training examples are unavailable, these methods can not be applied without other prior information about the sources. Examples of such prior information include "hummed" sounds mimicking the sources in the mixture [4], musical scores for music sources [5], or text transcriptions of the corresponding speech sources [6]. Among various spectral models used in the literature, non-negative matrix factorization (NMF) [2], [7], or its probabilistic formulation known as probabilistic latent component analysis (PLCA) [3], has been shown suitable for audio signals, and thus it is widely used in the ranges of state-of-the-art algorithms.

Recently the so-called user-guided approaches based on NMF have been proposed and shown to be very efficient in single channel case [8], [9]. These approaches allow the end-user to manually annotate information about activity of each sound source, *e.g.* if it is active or not, in temporal domain [8] or in time-frequency domains [9]. The annotated source activity information is then used, instead of isolated training data, to guide the separation process. This user annotation strategy has also been considered recently in interactive source separation approaches [10], [11]. Since the time-frequency annotation is difficult and very time consuming, even for experienced people, we consider the temporal annotation only in this paper where the user is simply asked to specify time segments where each source is active after hearing the mixture.

The temporal annotation-based audio source separation approach considered in this paper is motivated by the one from [8] where NMF is used as the source spectral model. We then introduce two complementary weighting strategies targeting at (i) re-weighting the data so as to enhance the importance of segments with less sources, and at (ii) re-weighting the data so as to re-equilibrate the impact of different segment types that may have different lengths. These weighting factors appear latter in the derived parameter estimation. The proposed weighted NMF approach was shown to outperform the baseline approach [8] and obtain comparable source separation performance with some other state-of-the-art approaches within the professionally produced music recordings (PPMR)[1] task of the fourth community-based Signal Separation Evaluation Campaign (SiSEC 2013). A video demonstration of the proposed approach integrated as a

[1]http://sisec.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings

plug-in into Audacity sound editing tool is available [2].

The rest of the paper is organized as follows. We first present the model and the baseline algorithm in Section II. We then describe the proposed weighting strategies in Section III followed by the derived parameter estimation and source reconstruction in Section IV. We discuss experimental evaluation in Section V, and we finally conclude in Section VI

II. Model and baseline approach

Let x_{fn} and $s_{j,fn}$ be the short-time Fourier transform (STFT) coefficients of the observed single-channel mixture signal and the j-th source signal, respectively, at frequency bin f and time frame n. The mixing model writes

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

where $f = 1, \ldots, F$, $n = 1, \ldots, N$, and J is the total number of sources. 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

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}, \tag{2}$$

where $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices of sizes $(F \times K)$ and $(K \times N)$, with $FK + KN << FN$, describing the source spectral patterns and the temporal activations, respectively. The parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ can be estimated by minimizing the following cost function:

$$C(\boldsymbol{\theta}) = \sum_{fn} d_{IS}\left(v_{fn}|\hat{v}_{fn}\right), \tag{3}$$

where $\hat{v}_{fn} = [\mathbf{WH}]_{fn}$ and $d_{IS}\left(x|y\right) = \frac{x}{y} - \log\frac{x}{y} - 1$ is Itakura-Saito (IS) divergence. The resulting multiplicative update (MU) rules for parameter estimation write:

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

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{WH})^{\cdot -2} \odot \mathbf{V}\right)\mathbf{H}^T}{\left(\mathbf{WH}\right)^{\cdot -1}\mathbf{H}^T} \tag{5}$$

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.

Many source separation algorithms, *e.g.* [3], first learn the spectral patterns $\mathbf{W}$ from some training data consisting of clean source examples, and then perform NMF decomposition on a test mixture, while keeping pre-learned $\mathbf{W}$ fixed. Temporal annotation based approach [8] introduces information about source activities via zeros in $\mathbf{H}$: *e.g.* if source 1 is represented by first two NMF components and is silent between frames 100 and 200, then one can set $\mathbf{H}(1, n) = \mathbf{H}(2, n) = 0$ for $n = 100, ..., 200$ while other entries are randomly initialized to positive values. The spectral patterns $\mathbf{W}$ are learned from all data without making any distinction between segments with clean sources and those with mixed sources. Thanks to the nature of the multiplicative update rule, the zero entries in $\mathbf{H}$ remain zero through iterations. In the next section, we present some weighting providing a continuum of strategies between these two, thus possibly allowing to choose a better intermediate strategy.

III. Proposed weighting strategies

We present two temporal weighting strategies to enhance the parameter estimation and explain how these strategies, being complementary, can be combined.

A. Weighting to enhance data purity

Given the temporal segmentation, the number of active sources and active components in each time frame n (denoted by $\mathrm{act\_src}_n$ and $\mathrm{act\_cmp}_n$, respectively) is known and the model estimation should rely more on frames with less active sources or less active components. Thus we define a weight factor as one of the following two options:

$$b_{fn}^{\mathrm{src}}(\lambda) = [1/\mathrm{act\_src}_n]^{\lambda}, \quad \forall f = 1, ..., F, \tag{6}$$

$$b_{fn}^{\mathrm{cmp}}(\lambda) = [1/\#(\mathrm{act\_cmp}_n)]^{\lambda}, \quad \forall f = 1, ..., F, \tag{7}$$

where $\lambda \geq 0$ is a fixed parameter. Note that $\lambda = 0$ corresponds to the case of no weighting as [8], while when $\lambda > 0$ higher weight value is assigned if less sources are active. An example of the value of $b_{fn}^{\mathrm{src}}(\lambda)$ corresponding to the number of active sources is represented on Fig. 1 (B).

B. Weighting to equilibrate segment types

In practice, different types of temporal segments may have very different lengths (*e.g.* "bass"-only segment is too long as compared to "piano"-only segment or "bass+vocals" segment is too long as compared to "piano+vocals" segment). Such a "dis-equilibrated" structure usually leads to over-fitting of the NMF model on a particular source active in a long segment, while modeling poorly other sources appeared only in short

[2]https://www.youtube.com/watch?v=EjpLKvphpMo&feature=youtu.be

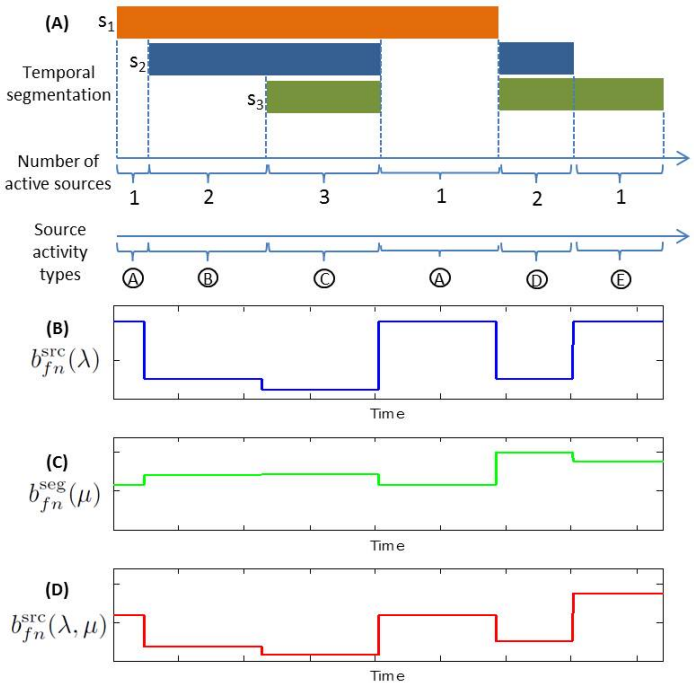

Fig. 1. Example of the proposed weighting strategies. **(A)**: temporal segmentation; **(B)**: weighting to enhance data purity $b_{fn}^{\mathrm{src}}(\lambda)$, $(\lambda = 2)$; **(C)**: weighting to equilibrate segment types $b_{fn}^{\mathrm{seg}}(\mu)$, $(\mu = 0.8)$; **(D)**: combined weighting $b_{fn}^{\mathrm{src}}(\lambda,\mu)$ $(\lambda = 2, \mu = 0.8)$.

segments. To overcome this issue we propose the following weighting

$$b_{fn}^{\mathrm{seg}}(\mu) = [1/\mathrm{seg\_len}_n]^{\mu}, \quad \forall f = 1, ..., F, \tag{8}$$

where $\mu \in [0, 1]$ is a fixed parameter, and $\mathrm{segm\_len}_n$ denotes the number of frames of the same segment type as the n-th frame (*e.g.* "piano+vocals" or "bass+vocals" segments). An example of the value of $b_{fn}^{\mathrm{seg}}(\mu)$ corresponding to the different segment types is represented on Fig. 1 (C).

C. Combined weighting

An unified weight combining the two above mentioned weighting strategies is finally defined as

$$b_{fn}^{\mathrm{src}}(\lambda,\mu) = b_{fn}^{\mathrm{src}}(\lambda)\, b_{fn}^{\mathrm{seg}}(\mu). \tag{9}$$

$$b_{lfn}^{\mathrm{cmp}}(\lambda,\mu) = b_{fn}^{\mathrm{cmp}}(\lambda)\, b_{fn}^{\mathrm{cmp}}(\mu). \tag{10}$$

Note that $b_{fn}^{\mathrm{src}}(\lambda,\mu)$ and $b_{fn}^{\mathrm{cmp}}(\lambda,\mu)$ vary over time frame n only and do not depend on frequency f since we are considering temporal annotation in this paper. Extensions to time-frequency weight are planed in our future work. An example of the value of $b_{fn}^{\mathrm{src}}(\lambda,\mu)$ is represented on Fig. 1 (C).

IV. Derived parameter estimation and signal reconstruction

To take into account the proposed weighting factors $b_{fn}(\lambda,\mu)$, NMF parameters are now estimated by minimizing the following cost function, which is slightly different from (3):

$$C_b(\boldsymbol{\theta}) = \sum_{fn} b_{fn}(\lambda,\mu) d_{IS}\left(v_{fn}|\hat{v}_{fn}\right). \tag{11}$$

The resulting multiplicative update (MU) rules for parameter estimation write:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\left((\mathbf{B}\odot\mathbf{WH})^{\cdot -2}\odot\mathbf{V}\right)}{\mathbf{W}^T(\mathbf{B}\odot\mathbf{WH})^{\cdot -1}} \tag{12}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left((\mathbf{B}\odot\mathbf{WH})^{\cdot -2}\odot\mathbf{V}\right)\mathbf{H}^T}{(\mathbf{B}\odot\mathbf{WH})^{\cdot -1}\mathbf{H}^T} \tag{13}$$

where $\mathbf{B} = \{b_{fn}(\lambda,\mu)\}_{fn}$ is a $F \times N$ weighting matrix.

Once the parameters are estimated, the STFT coefficients of the sources are obtained in the minimum mean square error (MMSE) sense by Wiener filtering as

$$\hat{s}_{j,fn} = \frac{\hat{v}_{j,fn}}{\hat{v}_{fn}} x_{fn}, \tag{14}$$

where $\hat{v}_{j,fn} = [\mathbf{W}_{(j)}\mathbf{H}_{(j)}]_{fn}$, $\mathbf{W}_{(j)}$ and $\mathbf{H}_{(j)}$ are subsets of $\mathbf{W}$ and $\mathbf{H}$, respectively, modeling the contribution of the j-th source. Finally, the time domain source signals are reconstructed via the inverse STFT.

V. Experiment

A. Data and evaluation criteria

We compared the source separation performance obtained by the proposed approach using weighted NMF with that obtained by the baseline method (when $\lambda = \mu = 0$) on the dev2 subset of the SiSEC2013's PPMRs task. This subset contains three stereo mixtures (18 - 25 second length and 44100 Hz sampling rate) to be separated together with the corresponding full 2 - 4 minute recordings. Note that the dev1 subset was not considered since it does not include full recordings, and thus the segmental information is very poor (almost all sources are active at the same time). Since the mixtures are multichannel here, we extended the weighted NMF algorithm to multichannel case exploiting spatial model, that is very similar to one presented in [12]. Since the benefit of spatial diversity and that of segmental temporal information are complementary and for sake of brevity and clarity, the detailed multichannel algorithm is not presented in this paper. However, readers can find more description in our longer report [13].

	Value	Average SDR (dB)					Average OPS (0 - 100)				
Weighting type	$\lambda \setminus \mu$	0	0.33	0.66	0.83	1	0	0.33	0.66	0.83	1
$b_{lfn}^{\mathrm{src}}(\lambda,\mu)$	0	*2.49*	2.20	1.96	2.36	2.42	*25.87*	22.60	26.09	24.88	24.76
	1	1.81	2.37	2.45	2.51	2.42	27.73	24.39	24.86	24.08	23.81
	3	2.42	2.74	2.59	2.09	3.09	24.62	25.93	25.16	23.86	24.06
	9	2.15	2.66	2.44	2.49	2.26	25.53	23.26	22.65	22.12	20.75
	27	1.99	2.22	1.85	1.97	2.24	21.17	20.08	19.84	19.75	20.12
$b_{lfn}^{\mathrm{cmp}}(\lambda,\mu)$	0	*2.49*	2.20	1.96	2.36	2.42	*25.87*	22.60	26.09	24.88	24.76
	1	2.15	2.36	2.54	2.38	2.59	26.86	24.80	26.78	25.68	24.30
	3	2.61	2.43	**3.10**	***3.00***	1.83	24.55	25.31	***28.02***	24.49	21.80
	9	2.81	1.79	2.13	2.57	2.34	**28.83**	25.55	21.27	24.36	22.69
	27	2.24	2.10	1.71	2.00	1.98	21.50	23.15	23.15	23.22	22.42

TABLE I
AVERAGE SDRS / OPSS FOR DIFFERENT WEIGHTING STRATEGIES ON THREE MIXTURES OF DEV2 DATASET OF SISEC 2013 PPMRS TASK.

The STFT is computed using a half-overlapping 93 ms (4096 samples) length sine window. The performance was assessed using the widely used criteria: signal-to-distortion ratio (SDR) [14] and overall perceptual score (OPS) [15], averaged for all sources and mixtures. These criteria have been widely used in recent international source separation evaluation campaign [1].

B. Oracle component grouping

Within the baseline user-guided source separation approach [8], a manual user intervention is needed in the first step (to create temporal segmentation) and may also be in the last step (to fine-tune NMF component grouping since neither temporal segmentation, nor spatial cues can completely disambiguate component - source assignments. For the faire comparison between different weighting strategies, we performed also the component grouping step in our experiment as follows. Given the available clean sources, the grouping is performed in an *oracle greedy* manner as inspired by [2]. First, source estimates are initialized by zero and estimated NMF components (reconstructed as in (14), but for NMF components indexed by $k = 1, \ldots, K$ instead of sources indexed by $j = 1, \ldots, J$) are ranged in order of decreasing energy. Then, while iterating over $k = 1, \ldots, K$, source estimates are updated by adding each time component to only one source so as the total mean squared error between the current source estimates and the clean sources is minimum.

C. Simulation results

The algorithm was run for both combined weighting strategies (9) and (10), for different parameter settings $(\lambda, \mu) \in \{0, 1, 3, 9, 27\} \times \{0, 0.33, 0.66, 0.83, 1\}$. The results are shown in table I. The best average results in terms of SDR are achieved with $b_{fn}^{\mathrm{cmp}}(\lambda = 3, \mu = 0.66)$, and those in terms of OPS with $b_{fn}^{\mathrm{cmp}}(\lambda = 9, \mu = 0)$. The corresponding configurations outperform the baseline non-weighted approach ($\lambda = \mu = 0$) by 0.6 dB for SDR and by 3 point for OPS.

For our SiSEC 2013 PPMRs task submission, we have chosen $b_{fn}^{\mathrm{cmp}}(\lambda = 3, \mu = 0.66)$. With this weighting we have also assessed the results on the same dev2 subset, but with manual component grouping instead of the oracle one. Average SDR and OPS are, respectively, 3.28 dB and 27.98, which is comparable to the oracle results (SDR $= 3.10$ dB and OPS $= 28.02$, see table I).

According to the SiSEC 2013 results[3] the performance of the proposed approach is comparable to that obtained by some state-of-the-art algorithms possibly using the same order of manual effort.

VI. CONCLUSION

In this paper, we have presented an user-guided audio source separation algorithm based on the weighted NMF. The proposed weighting strategies allow to efficiently exploiting the temporal annotation for learning the model parameters. Experiments with real-world mixtures used in international evaluation campaign confirm the effectiveness of the derived algorithm as compared to both the baseline and state-of-the-art approaches. Future research would be devoted to investigating time-frequency weights applied to time-frequency annotation, and applying these weighting strategies to other audio applications.

[3] http://www.onn.nii.ac.jp/sisec13/evaluation_result/MUS/testMUS2013.htm

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] 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.

[4] 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.

[5] S. Ewert, B. Pardo, M. Muller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[6] L. Le 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.

[7] 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.

[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] 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.

[11] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

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

[13] A. Ozerov, N. Q. K. Duong, and L. Chevallier, "Weighted nonnegative tensor factorization: on monotonicity of multiplicative update rules and application to user-guided audio source separation," Tech. Rep., Technicolor, Oct. 2013.

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

[15] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.