A Comparative Study of Example-guided Audio Source Separation Approaches Based On Nonnegative Matrix Factorization

Alexey Ozerov, Srđan Kitić, Patrick Pérez

2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

A COMPARATIVE STUDY OF EXAMPLE-GUIDED AUDIO SOURCE SEPARATION APPROACHES BASED ON NONNEGATIVE MATRIX FACTORIZATION

Alexey Ozerov, Srđan Kitić and Patrick Pérez

Technicolor
975 Avenue des Champs Blancs, 35576 Cesson-Sévigné, France

ABSTRACT

We consider example-guided audio source separation approaches, where the audio mixture to be separated is supplied with source examples that are assumed matching the sources in the mixture both in frequency and time. These approaches were successfully applied to the tasks such as source separation by humming, score-informed music source separation, and music source separation guided by covers. Most of proposed methods are based on nonnegative matrix factorization (NMF) and its variants, including methods using NMF models pre-trained from examples as an initialization of mixture NMF decomposition, methods using those models as hyperparameters of priors of mixture NMF decomposition, and methods using coupled NMF models. Moreover, those methods differ by the choice of the NMF divergence and the NMF prior. However, there is no systematic comparison of all these methods. In this work, we compare existing methods and some new variants on the score-informed and cover-guided source separation tasks.

Index Terms— Example-guided audio source separation, nonnegative matrix factorization, coupled nonnegative matrix factorization, comparative study

1. INTRODUCTION

Audio source separation remains still challenging [1], especially in the single-channel case. As such, one of the popular recent trends consists in turning from blind towards informed or guided source separation approaches [2], where some additional information about sources or mixing conditions is used so as to enhance the source separation quality. Many kinds of information were considered including music scores for music sources [3], text for speech sources [4], user-provided annotations [5, 6], a video corresponding to the sound mixture [7], audio-visual objects motion information [8], etc.

In this work we are interested in example-guided source separation [9–14], a particular sub-trend of guided approaches, where the additional information consists of source examples that are supposed to be close in some sense to the sources in the mixture, though do not coincide with them. More precisely, we consider here only the approaches where it is assumed that the examples match the sources both in frequency and time, thus excluding, e.g., the approaches supporting time-frequency deformations [15], pitch variations in case of speech sources [4], or the approaches where only spectral or temporal characteristics are matched [16]. The approaches we consider are suitable for the following tasks:

- *Humming-informed* source separation [9, 10], where the source examples are hummed by a user while listening to the mixture;
- *Score-informed* music source separation [11, 12], where the examples are synthesized from the corresponding music scores;
- *Cover-informed* music source separation [13, 14], where the examples are cover tracks played or sung by musicians.

To illustrate the main concepts, we show in Figure 1 some music sources, the corresponding mixture and source examples (synthesized from the scores as in [11,12]). Note that we here consider only the single-channel source separation, while for some methods, especially for those based on the Itakura-Saito (IS) divergence, extensions to multichannel case are quite straightforward, as it was done for example for a general source separation framework using multiple deformed references [17] extended to multichannel case in [15].

Most of NMF-based methods for example-guided audio source separation rely on the same global strategy schematized on Figure 1. First, the NMF models of the sources are estimated while trying to maintain a good compromise between fitting the models to both the examples and the mixture. Once the models have been estimated, the sources are estimated in turn by applying the corresponding NMF-driven Wiener filtering to the mixture. Estimating "good" NMF models that represent well the sources is the most critical step, and various strategies were proposed for that in different approaches. Those strategies are based on different training steps and criteria that are optimized in most cases using the multiplicative update (MU) rules [18, 19].

In [13], for cover-informed task, the NMF models are first learned from the examples, and then re-trained on the mixture. As such, the information from examples is only injected via models initialization for training from the mixture. Probabilistic latent component analysis (PLCA) modeling is considered in [9, 11] with application to humming-informed [9] and score-informed [11] tasks. Since PLCA modeling was shown equivalent (in terms of the criterion to be optimized) to the NMF with Kullback-Leibler (KL) divergence [20], these approaches fall into the scope considered here. In [9, 11] the models are first trained from the examples and then re-estimated from the mixture, while enforcing the models to be close to the example models within the estimation criterion, which is achieved by using example model parameters as hyper-parameters of some prior distributions. Within the probabilistic PLCA this may be also interpreted as a maximum *a posteriori* (MAP) adaptation. A similar approach was considered in [10] for humming-informed task, though with a different prior distribution and with a different model estimation algorithm (MU rules in [10] instead of the expectation-maximization (EM) algorithm as in [9, 11]). In this case the information from examples is injected into the final training from the mixture via both initialization of parameters and priors on them. Finally, coupled NMF modeling is considered in [14]

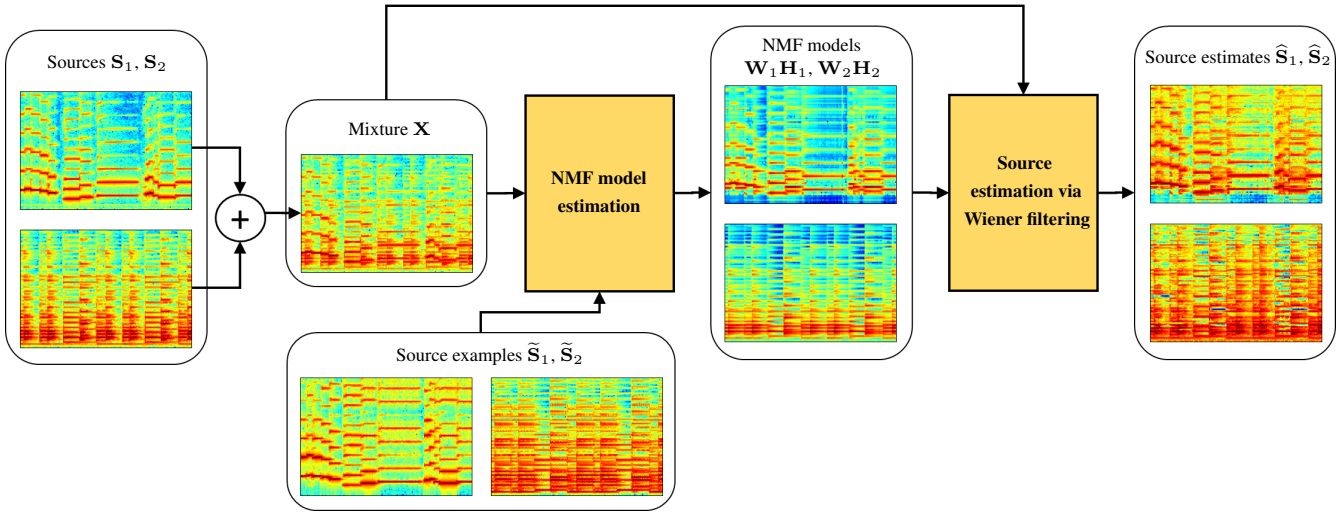

Fig. 1. A general scheme of example-guided audio source separation based on NMF modeling. Here it is illustrated on score-informed music source separation with examples synthesis [11, 12]. All the signals and models are represented by the corresponding spectrograms. The music sources (saxophone and piano) and the synthesized source examples are from TRIOS dataset [12].

for cover-informed task, where the NMF models are jointly trained from the examples and the mixture, while optimizing a composite criterion. As such, the information from examples (or the examples themselves) is directly used while learning from the mixture. Similar approaches, though applied in slightly different contexts, include those based on nonnegative matrix partial co-factorization (NMPCF) modeling [4, 21], where NMF model components [21] or NMF model factors (in case of a slightly more complex model) [4] are coupled only partially.

The state-of-the-art methods presented above differ not only by the model estimation strategies, but also by several factors including the choice of NMF divergence (e.g., KL divergence in [10] vs. IS divergence in [14]), by the choice of prior distribution in case of MAP-like model adaptation [9–11], and by the strategy of management of the example-guidance constraint. The latter factor means that the data (mixture) fit and the constraint (example) fit are often traded off via some penalty λ within the corresponding estimation criterion, and in some approaches [10, 11] the penalty λ is decreased over the algorithm iterations, and in other it is kept constant [14]. Also, as explained above, not all the approaches were applied to all tasks mentioned in the beginning of this section.

It is clear that there is a lack of comparison of different approaches on the same tasks, and the goal of this work is to fill in this gap. To do so we evaluate different existing approaches and their variants[1] on two informed source separation tasks, while varying several factors such as global estimation strategy, NMF divergence, prior distribution, etc. We consider score-informed source separation task and use TRIOS dataset [12] including 5 classic music mixtures for evaluation; and we consider cover-informed source separation task and use a dataset of 4 pop music mixtures introduced in [13], which we here refer to as *COVERS dataset*. To keep the evaluation as fair as possible, all the parameters are tuned via 5-fold or 4-fold cross-validation depending on the dataset.

The paper is organized as follows. Problem formulation and basic NMF modeling aspects are given in Section 2. Various model estimation strategies are presented in Section 3. Experimental protocol and simulations are given in Section 4 and conclusions are drawn in Section 5.

[1]Some variants of the approaches we evaluate might be considered as new, but in our opinion the novelty is incremental, since those variants are obtained by straightforward combination of existing bricks.

2. PROBLEM FORMULATION AND MODELING

2.1. Problem formulation

Let us consider a single-channel mixture of J sources

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

where x_{fn} and $s_{j,fn}$ are the short-time Fourier transform (STFT) coefficients[2] of the mixture and the j-th source, respectively, while $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote the frequency and time indices. It is assumed that there are also J source examples[3] of the same length available. Let $\tilde{s}_{j,fn}$ denote the corresponding STFT coefficients. Assuming all the signals rewritten in a matrix form as $\mathbf{X} = [x_{fn}]_{f,n}$, $\mathbf{S}_j = [s_{j,fn}]_{f,n}$ and $\widetilde{\mathbf{S}}_j = [\tilde{s}_{j,fn}]_{f,n}$, the problem consists in estimating the sources $\mathbf{S}_j$, given the mixture $\mathbf{X}$ and the examples $\widetilde{\mathbf{S}}_j$. The examples are used to guide the source separation process (see Fig. 1).

2.2. NMF modeling

Let $\mathbf{V}_x$, $\mathbf{V}_j$ and $\widetilde{\mathbf{V}}_j$ be the spectrograms or the power spectrograms (i.e., $\mathbf{V}_x = |\mathbf{X}|$ or $\mathbf{V}_x = |\mathbf{X}|^{\cdot 2}$ and the same for other matrices[4]) of the mixture, the sources and the examples, respectively. Within

[2]Within this paper all the signals are introduced directly in the STFT domain, assuming that they can be always computed from the time signals with an appropriate STFT, and that the respective time signals can be always reconstructed by a corresponding inverse STFT.

[3]In a more general formulation it might be supposed that not all J sources are supplied with the corresponding examples [4, 17], though for the sake of simplicity we here assume that there are always J examples.

[4]Within this paper the absolute value and the power operations are applied to matrices element-wise.

the NMF-based approaches considered here the nonnegative (power) spectrograms of the latent sources are assumed approximated as

$$\mathbf{V}_j \approx \mathbf{W}_j \mathbf{H}_j, \tag{2}$$

where $\mathbf{W}_j$ and $\mathbf{H}_j$ are both nonnegative matrices of sizes $F \times K_s$ and $K_s \times N$, respectively; and the model order (also called the number of NMF components) K_s is usually chosen smaller than both F and N. Here K_s is chosen the same for each source, while it is also possible to adapt it to each source [22] by finding an optimal $K_{s,j}$ per source among a fixed total number of components $K = \sum_{j=1}^{J} K_{s,j}$.

Assuming $\mathbf{V}_j$ known, NMF model parameters

$$\boldsymbol{\theta}_j = \{\mathbf{W}_j, \mathbf{H}_j\} \tag{3}$$

are usually estimated by minimizing some measure of fit between $\mathbf{V}_j$ and $\mathbf{W}_j\mathbf{H}_j$ as

$$\boldsymbol{\theta}_j = \arg\min_{\boldsymbol{\theta}'_j} D(\mathbf{V}_j \| \mathbf{W}'_j \mathbf{H}'_j), \tag{4}$$

where $D(\cdot\|\cdot)$ is a divergence computed as a sum over time-frequency indices of a scalar divergence $d(\cdot|\cdot)$ as

$$D(\mathbf{A}\|\mathbf{B}) = \sum\nolimits_{f,n=1}^{F,N} d(a_{fn}|b_{fn}), \tag{5}$$

$\mathbf{A} = [a_{fn}]_{f,n}$ and $\mathbf{B} = [b_{fn}]_{f,n}$ being some $F \times N$ nonnegative matrices. The most popular scalar divergences we will consider here include the KL divergence [18]

$$d_{\mathrm{KL}}(a|b) = a \log \frac{a}{b} - a + b, \tag{6}$$

the IS divergence [19]

$$d_{\mathrm{IS}}(a|b) = \frac{a}{b} - \log \frac{a}{b} - 1, \tag{7}$$

and the Euclidean (EUC) distance [18]

$$d_{\mathrm{EUC}}(a|b) = |a - b|^2 . \tag{8}$$

However, EUC distance was not found as efficient as two other divergences for audio-related applications. Optimization in (4) is usually achieved by applying multiplicative update (MU) rules [18, 19].

However, $\mathbf{V}_j$ are unknown, and the main problem consists in finding sufficiently good approximations (2), while using only the mixture and the examples.

2.3. Source estimation

Once the NMF decompositions (2) are obtained, the source estimates $\widehat{\mathbf{S}}_j$ are usually computed by Wiener filtering as follows

$$\widehat{\mathbf{S}}_j = \frac{\mathbf{W}_j \mathbf{H}_j}{\sum_{j=1}^{J} \mathbf{W}_j \mathbf{H}_j} \odot \mathbf{X}, \tag{9}$$

where the matrix division is element-vise and $\odot$ denotes element-vise matrix multiplication.

3. NMF MODEL ESTIMATION STRATEGIES

Let source NMF models (3) be concatenated into one mixture NMF model with JK_s components as

$$\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_J], \qquad \mathbf{H} = [\mathbf{H}_1^T, \ldots, \mathbf{H}_J^T]^T, \tag{10}$$

which is parametrized as:

$$\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}. \tag{11}$$

We consider the following four model estimation strategies.

3.1. Strategy #1 - Retrained NMF

This strategy was considered in [13] and consists in following steps:

1. Example models $\tilde{\boldsymbol{\theta}}_j = \{\widetilde{\mathbf{W}}_j, \widetilde{\mathbf{H}}_j\}$ are initialized with random nonnegative values and estimated optimizing the following criterion (similar to criterion (4)):

$$\tilde{\boldsymbol{\theta}}_j = \arg\min_{\tilde{\boldsymbol{\theta}}'_j} D(\widetilde{\mathbf{V}}_j \| \widetilde{\mathbf{W}}'_j \widetilde{\mathbf{H}}'_j). \tag{12}$$

2. Mixture model $\boldsymbol{\theta}$ (11) is initialized with example models $\tilde{\boldsymbol{\theta}}_j$ and retrained from the mixture optimizing the following criterion:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}'} D(\mathbf{V}_x \| \mathbf{W}' \mathbf{H}'). \tag{13}$$

3.2. Strategy #2 - Prior NMF

A so-called *prior NMF* strategy adopted in [9–11] consists in following steps:

1. Example models $\tilde{\boldsymbol{\theta}}_j$ are initialized with random nonnegative values and estimated optimizing criterion (12), as for retrained NMF strategy.
2. Mixture model $\boldsymbol{\theta}$ (11) is initialized with example models $\tilde{\boldsymbol{\theta}}_j$ and retrained from the mixture optimizing the following constrained criterion:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}'} D(\mathbf{V}_x \| \mathbf{W}' \mathbf{H}') + \lambda \left[\Psi_W(\mathbf{W}' \| \widetilde{\mathbf{W}}) + \Psi_H(\mathbf{H}' \| \widetilde{\mathbf{H}}) \right], \tag{14}$$

with $\lambda \geq 0$ being a penalty factor, and

$$\Psi_W(\mathbf{W}\|\widetilde{\mathbf{W}}) = \frac{N}{K} \sum\nolimits_{k,f=1}^{K,F} \psi(w_{fk}|\tilde{w}_{fk}), \tag{15}$$

$$\Psi_H(\mathbf{H}\|\widetilde{\mathbf{H}}) = \frac{F}{K} \sum\nolimits_{k,n=1}^{K,N} \psi(h_{kn}|\tilde{h}_{kn}), \tag{16}$$

where $\psi(a|\tilde{a})$ is a measure of fit between a parameter $\tilde{a}$ of the example models and the corresponding parameter a of the mixture model.

Normalization factors N/K and F/K in (15) and (16) serve to compensate for the fact that the summation in $D(\mathbf{V}_x\|\mathbf{WH})$ includes FN terms (see (5)), while the summations in $\Psi_W(\mathbf{W}\|\widetilde{\mathbf{W}})$ and $\Psi_H(\mathbf{H}\|\widetilde{\mathbf{H}})$ are over KF and KN terms, respectively.

In probabilistic settings, as considered in [9–11], $D(\mathbf{V}_x\|\mathbf{WH})$ represents negative log-likelihood, while $\Psi_W(\mathbf{W}\|\widetilde{\mathbf{W}})$ and $\Psi_H(\mathbf{H}\|\widetilde{\mathbf{H}})$ represent negative log-priors, with $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{H}}$ being some hyperparameters. In that case, criterion (14) corresponds to a MAP criterion.

In [9, 11], $\psi(a|\tilde{a})$ is defined via a Dirichlet prior, which leads to:[5]

$$\psi_{\mathrm{Dirichlet}}(a|\tilde{a}) = -\log \left[a^{\tilde{a}} \right]. \tag{17}$$

However, we do not claim here to reproduce the methods from [9, 11]. Indeed, due to slightly different PLCA modeling a and $\tilde{a}$ are conditional probabilities in [9, 11], thus they sum to 1 over k, while we do not assume such a constraint here. Moreover, we here use MU rules instead of an EM algorithm as in [9, 11], which may also lead

[5] Note that in the expressions for distributions in (17) and (18) constant (i.e., independent on a) multiplicative factors and powers are omitted, since constant multiplicative factors do not have any influence on the optimization in (14), and constant powers are assumed to be absorbed into λ in (14).

to a different result. As such, we only claim applying prior (17) to the NMF model in our evaluation.

In [10], a Gamma distribution is considered, which leads to

$$\psi_{\mathrm{Gamma}}(a|\tilde{a}) = -\log\left[a^{\alpha-1}\tilde{a}^{-\alpha}\exp(a/\tilde{a})\right], \quad (18)$$

and it is moreover assumed that $\alpha = 1$.

Finally, we here consider three new measures of fit, notably

$$\psi_{\mathrm{KL}}(a|\tilde{a}) = d_{\mathrm{KL}}(a|\tilde{a}), \quad (19)$$
$$\psi_{\mathrm{IS}}(a|\tilde{a}) = d_{\mathrm{IS}}(a|\tilde{a}), \quad \text{and} \quad (20)$$
$$\psi_{\mathrm{EUC}}(a|\tilde{a}) = d_{\mathrm{EUC}}(a|\tilde{a}), \quad (21)$$

where the corresponding divergences are defined in (6), (7) and (8).

3.3. Strategy #3 - Coupled NMF

This strategy [4, 14] consists just in one step, where the NMF models are estimated jointly (this is why it is also called *joint NMF* in [14]) from both the mixture and the examples by optimizing the following criterion:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}'} D(\mathbf{V}_x\|\mathbf{W}'\mathbf{H}') + \lambda\sum\nolimits_{j=1}^{J} D(\widetilde{\mathbf{V}}_j\|\mathbf{W}'_j\mathbf{H}'_j). \quad (22)$$

As usual, the NMF model parameters are initialized with random nonnegative values.

3.4. Strategy #4 - Supervised NMF

For a completeness of the picture we are considering as well a so-called *supervised NMF* [23] strategy, where the example NMF models are not adjusted on the mixture at all, and used as they are to preform separation. This is simply achieved by discarding the second step in the retrained NMF or the prior NMF strategy, i.e., the models are just learned from the examples by optimizing criterion (12).

3.5. Model estimation algorithms

All the criteria within this work are optimized via MU rules derived using common heuristics as in [18, 19, 24].

3.6. Discussion

While model estimation strategies #1 to #3 are all based on a trade-off between fitting the NMF models to both the examples and the mixture, the supervised NMF strategy #4 estimates the models from the examples only.

It is also worth to mention that the retrained NMF and the supervised NMF strategies are both limit cases of the prior NMF strategy. Indeed, it is easy to see that with $\lambda = 0$ in (14) the prior NMF strategy reduces to the the retrained NMF strategy, and with λ in (14) having a very high value it reduces to the supervised NMF strategy. As for the coupled NMF strategy, it reduces to the supervised NMF strategy with λ having a very high value. As such, we believe that more sophisticated strategies #2 and #3 (prior and coupled NMF) should lead to better results than simpler strategies #1 and #4 (retrained and supervised NMF), which has been already partially observed in [14].

4. EXPERIMENTS

4.1. Data

For the score-informed task we use the TRIOS dataset [12]. It includes 5 mixtures of length from 18 to 53 seconds of classic music performances, with 4 mixtures of 3 sources and one mixture of 5 sources. The dataset includes source examples as well, which are already synthesized by a MIDI synthesized from the corresponding scores and temporally aligned with the respective mixtures. All the audio signals are sampled at 44100 Hz.

For the cover-informed task we use the COVERS dataset introduced in [13] and further used in [14]. The dataset consists of 4 professionally produced pop music recordings of length from about 2 to 5 minutes. Each recording is a mixture of 4 to 6 tracks (sources) to be separated. Moreover, the dataset includes cover tracks played by musicians. The cover tracks are temporally aligned to the respective mixtures. All the audio signals are sampled at 44100 Hz. Following [13] we have resampled all the signals to 32000 Hz and have retained for the experiments one 30 second expert per mixture, as it is done in a preview of the dataset that can be found at [6].

4.2. Evaluation metrics

Following [13] and [14] we measure the quality of the estimated sources in terms of the signal-to-distortion ratio improvement (SDRI) that is the difference between the output signal-to-distortion ratio (SDR) computed as proposed in [25] and the input SDR. The input SDR is defined as the power ratio between the source to be estimated and the mixture to be separated. A similar metric called normalized SDR (NSDR) was proposed in [26].

To obtain an average SDRI over several mixtures, it is first averaged over the sources in each mixture and then over all the mixtures.

4.3. Parameters

The STFTs are computed with half-overlapping sine windows of length 2048, *i.e.* 46 ms, for the TRIOS dataset, and 64 ms, for the COVERS dataset. Each MU rules algorithm is run for 50 iterations. Following usual practice [18, 19] we used spectrograms (i.e., $\mathbf{V}_x = |\mathbf{X}|$, etc.) for KL-NMF and EUC-NMF decompositions and power spectrograms (i.e., $\mathbf{V}_x = |\mathbf{X}|^{\cdot 2}$, etc.) for IS-NMF decompositions.

For model estimation strategies including a penalty factor λ (strategies #2 and #3) we consider the following two cases:

- *Fixed*: $\lambda = \lambda_0$ is kept constant over the algorithm iterations as in [14].
- *Decreased*: λ is linearly decreased from λ_0 to 0 over the algorithm iterations as in [10, 11].

4.4. Evaluation protocol

Simulations were carried using L-fold cross-validation with $L = 5$ for TRIOS dataset and $L = 4$ for COVERS dataset. Parameters K_s and λ_0 are varied over the following grid

$K_s = \{5, 10, 15, 20, 30, 40, 50, 70, 100, 150, 200, 300\}$,

$\lambda_0 = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$,

while λ_0 is not varied for strategies #1 and #4.

[6] http://www.gipsa-lab.grenoble-inp.fr/~laurent.girin/demo/ismir2012.html

Score-informed task – TRIOS dataset

Divergence \ Strategy	Strategy #1 Retrained	Strategy #2 - Prior					Strategy #3 Coupled	Strategy #4 Supervised
		Dirichlet	Gamma	KL prior	IS prior	EUC prior		
KL	13.94 *(100)*	13.91 / 13.93 *(100 / 100)*	13.47 / 13.40 *(56 / 70)*	13.93 / 13.94 *(100 / 100)*	14.37 / 14.42 *(300 / 300)*	13.94 / 13.94 *(100 / 100)*	13.94 / 13.94 *(100 / 100)*	11.76 *(34)*
IS	12.66 *(180)*	12.37 / 12.55 *(180 / 180)*	13.64 / 14.04 *(134 / 150)*	13.52 / 13.78 *(270 / 300)*	14.71 / 14.23 *(150 / 180)*	12.33 / 12.38 *(180 / 180)*	13.76 / 13.68 *(134 / 134)*	11.73 *(134)*
EUC	13.11 *(270)*	13.24 / 13.21 *(300 / 300)*	13.18 / 13.15 *(260 / 230)*	13.29 / 13.31 *(300 / 300)*	13.81 / 13.8 *(100 / 88)*	13.15 / 13.16 *(300 / 300)*	13.11 / 13.11 *(270 / 270)*	11.74 *(100)*

Cover-informed task – COVERS dataset

KL	10.31 *(188)*	10.14 / 10.4 *(200 / 275)*	9.8 / 10 *(80 / 89)*	10.86 / 11.19 *(163 / 225)*	10.89 / 10.94 *(19 / 9)*	10.55 / 10.64 *(225 / 250)*	10.99 / 11.2 *(15 / 17)*	10.6 *(15)*
IS	9.17 *(58)*	9.17 / 9.27 *(58 / 58)*	10.2 / 10.25 *(56 / 43)*	10.24 / 10.31 *(105 / 75)*	10.74 / 10.2 *(70 / 50)*	10.3 / 10.18 *(85 / 83)*	10.73 / 10.61 *(50 / 30)*	10.11 *(28)*
EUC	10.62 *(200)*	10.42 / 10.48 *(200 / 200)*	10.61 / 10.62 *(200 / 200)*	10.33 / 10.3 *(160 / 185)*	10.64 / 10.91 *(14 / 20)*	10.98 / 10.95 *(187 / 200)*	10.38 / 10.5 *(107 / 62)*	10.57 *(17)*

Table 1. Source separation performance of each method measured in terms of the average SDRI (in dB). The average optimal K_s value is reported as well below in parentheses. As for the methods involving penalty λ, the values at the left correspond to the fixed λ, while the values at the right correspond to the decreased λ.

4.5. Simulation results

The results are reported in Table 1, where we have also included the average optimal K_s value in parentheses, since this value impacts the computational load of the corresponding approach. First, it should be noted that there is not much variation of performance between different approaches: results vary within 3 dB for TRIOS dataset and within 2.5 dB for COVERS dataset. Also, there is not so much difference between keeping penalty λ constant and decreasing it, though for the COVERS dataset decreasing λ leads in average to slightly better results. As for the NMF divergence, it could be noted that for both datasets the KL divergence performs in average better than the IS divergence. Surprisingly and contrary to what is usually reported for other NMF-based audio processing methods, the EUC distance performs quite well in this constrained setting. As expected (see discussion in Section 3.6), more sophisticated strategies #2 and #3 outperform in most cases strategies #1 and #4. The best SDRI for TRIOS dataset is obtained by prior NMF strategy with IS divergence and IS prior, and for COVERS dataset by prior NMF strategy with KL divergence and KL prior and by coupled KL-NMF. Those methods are slightly new variants of the existing methods. Assuming our experimental settings for COVERS dataset reproduce those from [14], our best SDRI is on par or even slightly better than the best SDRI reported in [14].

It is also interesting to note that while for TRIOS dataset supervised NMF strategy performs worth than all other strategies, for COVERS dataset the result of the supervised NMF with KL divergence is not so bad, the corresponding SDRI is only 0.5 dB below the best SDRI. This is an important observation, since the supervised NMF strategy is much simpler than all others, it does not require any model retraining or adaptation on the mixture.

The average optimal number of components K_s varies quite drastically depending on the method and it also often saturates to the maximal $K_s = 300$ we tested. One can only note that the supervised strategy requires in most cases less components, possibly because it might suffer from data overfitting, which is not the case for the other methods.

Finally, we have also measured source separation performance obtained without any modeling by directly using the power spectrograms of either examples or the sources. In the latter case it gives just an *oracle* (a kind of upper bound) performance, since the sources are not known. This was achieved by simply replacing $\mathbf{W}_j\mathbf{H}_j$ in the Wiener filtering (9) with either $|\widetilde{\mathbf{S}}_j|^{\cdot 2}$ or $|\mathbf{S}_j|^{\cdot 2}$. This lead to the average performances of 12.36 dB and 24.09 dB for TRIOS dataset and of 3.49 dB and 15.82 dB for COVERS dataset. These results indicate that the NMF modeling is important, since the best results reported in Table 1 are substantially higher than 12.36 dB and 3.49 dB obtained by a direct use of the example power spectrograms, especially for the COVERS dataset. This indicates also that the COVERS dataset is more difficult than the TRIOS one, i.e., there is more mismatch between the sources and the source examples (a direct use of examples as proxies for the sources leads to a quite bad separation). This was confirmed by an informal listening to the sources and examples; and by visual comparison of their spectrograms. Finally, the oracle performances of 24.09 dB and 15.82 dB indicate that there is still a room for improvement to go beyond the above investigated NMF modeling strategies.

5. CONCLUSIONS

In this work we carried out an experimental comparison of example-guided audio source separation approaches, where the audio mixture is supplied with source examples. We compared several existing NMF-based strategies and slightly new variants of them on score-informed and cover-informed music source separation tasks using TRIOS and COVERS datasets, respectively. We have found that the best results on both datasets were achieved by prior NMF strategy with KL divergence and IS or KL prior, which are new variants of existing methods we considered.

Further research on this topic should most probably focus on introducing knowledge-based constraints or deformations within the NMF modeling, as it was already started in [4, 15]. Other possible directions are to consider more sophisticated NMF models (e.g., convolutive NMF [27]), to reconsider NMF estimation procedures (e.g., discriminative training [28]), or to develop deep learning-based methods [29].

6. REFERENCES

[1] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H.i Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[2] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

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

[4] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[5] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multi-channel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing*, Prague, Czech Republic, May 2011, pp. 257 – 260.

[6] N. J. Bryan, G. J. Mysore, and G. Wang, "ISSE: An interactive source separation editor," in *the Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Toronto, Canada, April 2014.

[7] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," in *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing*, Philadelphia, USA, 2005, pp. 425–428.

[8] S. Parekh, S. Essid, A. Ozerov, N. Duong, P. Perez, and G. Richard, "Motion informed audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, 2017.

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

[10] D. FitzGerald, "User assisted source separation using non-negative matrix factorisation," in *22nd IET Irish Signals and Systems Conference*, Dublin, 2011.

[11] J. Ganseman, G. J. Mysore, J.S. Abel, and P. Scheunders, "Source separation by score synthesis," in *Proc. Int. Computer Music Conference (ICMC)*, New York, NY, June 2010, pp. 462–465.

[12] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Int. Conf. on Acoustics, speech, and signal processing*, 2013.

[13] T. Gerber, M. Dutasta, L. Girin, and C. Févotte, "Professionally-produced music separation guided by covers," in *International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, Oct. 2012.

[14] N. Souviraá-Labastie, E. Vincent, and F. Bimbot, "Music separation guided by cover tracks: designing the joint NMF model," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) 2015*, 2015.

[15] N. Souviraá-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, pp. 1775–1787, 2015.

[16] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation - A novel user-friendly framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 261–272, 2017.

[17] N. Souviraá-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Audio source separation using multiple deformed references," in *European Signal Processing Conference (EUSIPCO)*, 2014.

[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural and Information Processing Systems 13*, 2001, pp. 556–562.

[19] 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.

[20] E. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in *Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIRŠ05)*, New York, NY, USA, 2005, pp. 601–602.

[21] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[22] C. Bilen, A. Ozerov, and P. Pérez, "Automatic allocation of NTF components for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*, Shanghai, China, Mar. 2016.

[23] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 414–421.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.

[25] 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.

[26] 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.

[27] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs.," in *Fifth International Conference on Independent Component Analysis*, Granada, Spain, Sep. 2004, pp. 494–499.

[28] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation.," in *INTERSPEECH*, 2014, pp. 865–869.

[29] M. Miron, J. Janer, and E. Gómez, "Monaural score-informed source separation for classical music using convolutional neural networks," in *18th International Society for Music Information Retrieval Conference*, Suzhou, China, Oct. 2017.