HAL
archives-ouvertes.fr

Automatic Allocation of NTF Components for User-Guided Audio Source Separation

Cagdas Bilen, Alexey Ozerov, Patrick Pérez

▶ **To cite this version:**

Cagdas Bilen, Alexey Ozerov, Patrick Pérez. Automatic Allocation of NTF Components for User-Guided Audio Source Separation. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16), Mar 2016, Shanghai, China. 2016. <hal-01259430>

HAL Id: hal-01259430
https://hal.inria.fr/hal-01259430

Submitted on 20 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AUTOMATIC ALLOCATION OF NTF COMPONENTS FOR USER-GUIDED AUDIO SOURCE SEPARATION

Çağdaş Bilen, Alexey Ozerov and Patrick Pérez

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{cagdas.bilen, alexey.ozerov, patrick.perez}@technicolor.com

ABSTRACT

Nonnegative matrix or tensor factorization is a very popular approach for audio source separation. One important problem in nonnegative tensor factorization (NTF) in the context of user-guided audio source separation is the necessity to manually assign the NTF components to audio sources in order to be able to enforce prior information on the sources during the estimation process. In this paper, two new approaches to NTF based source separation are proposed, which do not require any manual component assignment to the sources, but estimate the underlying assignment automatically. Both algorithms use the prior information on the source samples in the estimation process along with either a limit on the minimum number of components each source uses or with a restriction that each component is used by sparse number of sources. The proposed methods are shown to outperform the classic approach with a manual distribution of the components equally among the sources.

***Index Terms*—** Nonnegative matrix factorization, Itakura-Saito divergence, generalized expectation-maximization, source separation

1. INTRODUCTION

Audio source separation remains still very challenging, especially in the single-channel case and for reverberant mixtures [1]. Moreover, the resulting source separation performance depends greatly on the amount of prior information about the sources and/or the mixing process that can be incorporated within the corresponding source separation algorithm [2]. One of new trends in audio source separation, referred as *user-guided* or *user-assisted*, consists in providing such prior information directly by a user [3–9], and many of those approaches rely on time-frequency annotations of the spectrograms [10–18].

Early approaches [10–12,16] are based on time annotations only, i.e., a user specifies (e.g., via a dedicated graphical user interface) which source is active at which moment. Then, the time annotations were extended to more general and flexible time-frequency annotations [13, 17, 18]. Finally, interactive frameworks [14, 15], where user has a possibility of gradually completing and correcting time-frequency annotations, were proposed as well. In addition, within an interactive framework Duong *et al.* [15] proposed a method allowing dealing with early stage separation errors through uncertainty propagation principle and Jeong and Lee [18] proposed a method allowing dealing with user annotation errors through a sparsity-inducing ℓ_1-norm penalty.

As for the modeling, most of the approaches are based on popular nonnegative matrix factorization (NMF) or nonnegative tensor factorization (NTF) approximations [10–16] and many on NMF/NTF with Itakura-Saito (IS) divergence [12, 13, 15, 16]. The use of NMF/NTF with Itakura-Saito (IS) divergence [19] (as compared to other popular loss functions such as Euclidean distance or Kullback-Leibler divergence [20]) is attractive for several reasons. First, its scale-invariance property [19] makes it more suitable for modeling audio spectrograms. Second, its probabilistic Gaussian formulation [19] facilitates many extensions such as multichannel formulations [2, 12, 21] and separation errors management via uncertainty propagation [15]. As such, we consider in this work NMF/NTF modeling with IS divergence. Finally, there are also methods that are based on nuclear norm as a low-rank inducing penalty [17, 18], which usually leads to convex optimization problems.

All the above-mentioned NMF/NTF-based methods [10–16] suffer from the following important problem. The number of NMF/NTF components (i.e., rank-1 matrices or tensors) affected to each source must be defined in advance instead of being learned given a total budget of K components. For example, one needs specifying in advance: 4 components for "bass guitar", 10 components for "piano", and 6 components for "drums" (see, e.g., [12]).[1] This problem leads to the following potential disadvantages. First, the user spends more time by choosing a suitable number of components K_j ($j = 1, \ldots, J$, and J is the total number of sources) instead of choosing just one total number of components K. Second, a suitable number of components chosen by user may be quite different from the one that would lead to the best source separation performance. This problem possibly arises from the fact that the modeling and the management of source activity constraints are not well separated. Indeed, a usual approach to manage time constraints is to set corresponding temporal activations to zero [10–12, 16], and thus one needs to say in advance which component is affected to which source. However, in case of time-frequency annotations the latter trick does not work, and these constraints are usually managed via heuristic penalizations of the corresponding cost functions [13–15]. However, such penalizations need as well the information about allocation of the components among the sources.

In order to overcome the aforementioned limitations of the existing NMF/NTF-based methods we propose in this work a new method based on time-frequency annotations and the NTF model with IS divergence as in [12]. The main novelty of our proposal is that we change the way the time-frequency annotations are taken into

[1] Choosing equal number of NMF/NTF components per source is usually suboptimal, since some sources have more spectral diversity than others, and thus need more components to be well represented. For example, it is evident that "piano" needs more components than "bass guitar", simply because a piano may produce about twice as many possible notes as a bass guitar.

account for model and sources estimation. Instead of using heuristic cost function penalizations as in [13–15] we propose estimating the NTF model from all available observations in the maximum likelihood (ML) or constrained ML sense under the corresponding Gaussian modeling (as in [19]). We also slightly change the way the sources are estimated given the model. This is achieved via Wiener filtering as in [13–15] with the only difference that the filter is applied not only to the mixture, but to all available observations (mixture and partial observations of sources).

The main advantage of the proposed approach over the state of the art is that it does not require the number of NTF components per source K_j to be specified. One only needs specifying the total budget of components K that it then automatically allocated between sources during model estimation. Among the two variations of the proposed approach, the first one assigns a minimum number of components to each source and automatically estimates the assignment of the remaining components. The second variation assumes a sparse structure on the NTF coefficients so that the components contribute to a *sparse* number of sources to better model the independent behavior of the sources. There are also few secondary potential advantages to the proposed method. First, as it was already mentioned, the optimization criterion is not designed from some heuristic considerations, but it is the ML or a constraint ML criterion for the given model and observations. Second, as it will be explained more in detail below, the time-frequency annotation constraints are taken into account on both steps of estimation of the model and estimation of the sources, while in the state-of-the-art [13–15] these constraints are only taken into account in the model estimation stage. We evaluate our approach in case of temporal annotations on a dataset of music mixtures and compare it with the strategy of uniformly allocating the components.

The rest of the paper is organized as follows. The problem to be solved is introduced and the proposed approach is described in Section 2. Section 3 is devoted to experiments and some conclusions are drawn in Section 4.

2. PROPOSED APPROACH

2.1. Problem Definition

Let us consider a single channel mixture composed of J sources with the mixing equation

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

in which x_{fn} is the Short-Time Fourier Transform (STFT) coefficient of the mixture at frequency bin $f \in [\![1, F]\!]$ and time window $n \in [\![1, N]\!]$ and s_{jfn} is the STFT coefficient of the source j for the same time-frequency coordinate. A subset, $\mathbf{\Omega} \subset [\![1, J]\!] \times [\![1, F]\!] \times [\![1, N]\!]$, of the time-frequency samples of the sources are assumed to be known up to an additive noise factor b_{jfn} such that

$$y_{jfn} = s_{jfn} + b_{jfn}, \quad \text{for all} \quad (j, f, n) \in \mathbf{\Omega}. \tag{2}$$

The problem that is considered in this paper is the problem of estimating the source signals, $\{s_{jfn}\}_{j,f,n}$, given the mixture, $\{x_{fn}\}_{f,n}$ and a subset of the measured source samples, $\{y_{jfn}\}_{(j,f,n)\in\mathbf{\Omega}}$. In practice the observed samples can be set as annotated silent periods of the sources.[2] Hence for the rest of this paper we shall assume that the support of measurements, $\mathbf{\Omega}$, indicates the time frequency points for which the corresponding source is known to be silent, i.e. $y_{jfn} = 0, \forall (j, f, n) \in \mathbf{\Omega}$.

[2]The algorithms presented in this paper are formulated directly in STFT domain assuming the partial source observations are formulated in this domain, e.g., via some user annotations as in [13–15]. However a formulation with respect to time domain samples is also possible as presented in [22].

2.2. Model Assumptions

The measurement noise, b_{jfn}, and the source samples, s_{jfn}, are assumed to be independently Gaussian distributed with variances $\sigma^2_{b,jfn}$ and v_{jfn} respectively such that

$$b_{jfn} \sim \mathcal{N}_c\left(0, \sigma^2_{b,jfn}\right), \quad s_{jfn} \sim \mathcal{N}_c(0, v_{jfn}). \tag{3}$$

The tensor of the source power spectrum is modeled to be low rank such that

$$v_{jfn} = \sum\nolimits_{k=1}^{K} q_{jk} w_{fk} h_{nk}, \tag{4}$$

where the number of components, K, is sufficiently small. This low rank tensor model is defined by the nonnegative matrices $\mathbf{Q} = [q_{jk}]_{j,k}$, $\mathbf{W} = [w_{fk}]_{f,k}$ and $\mathbf{H} = [h_{nk}]_{n,k}$ being, respectively, $J \times K$, $F \times K$ and $N \times K$. The NTF model parameters are represented by $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{W}, \mathbf{H}\}$. Similar to NMF, the matrix $\mathbf{H}$ and $\mathbf{W}$ represent the power distribution of the K components among the time and frequency bins respectively. The matrix $\mathbf{Q}$ represents the power distribution of the K components among the sources.

2.3. NTF Model Constraints

Assuming the tensor $\mathbf{P} = [p_{jfn}]_{j,f,n}$ of the power spectra is known (with $p_{jfn} = |s_{jfn}|^2$), NTF model parameters can be estimated using the multiplicative update (MU) rules minimizing the Itakura-Saito (IS) divergence [19] between the given 3-valence tensor of source power spectra, $\mathbf{P}$, and the 3-valence tensor of the NTF model approximation, $\mathbf{V} = [v_{jfn}]_{j,f,n}$, defined as

$$D_{IS}(\mathbf{P}\|\mathbf{V}) = \sum\nolimits_{j,f,n} d_{IS}(p_{jfn}\|v_{jfn}), \tag{5}$$

where $d_{IS}(x\|y) = x/y - \log(x/y) - 1$ is the IS divergence. The motivation to use the IS divergence instead of some other possible divergences (such as Kullback-Leibler divergence) is that the solution minimizing the IS divergence is shown to be equivalent to estimating in the ML sense [19]. The following simple multiplicative update (MU) rules are derived (as in [12]) to estimate the model parameters, $\mathbf{Q}, \mathbf{W}, \mathbf{H}$, to minimize the IS divergence:

$$q_{jk} \leftarrow q_{jk} \left(\frac{\sum_{f,n} w_{fk} h_{nk} p_{jfn} v_{jfn}^{-2}}{\sum_{f,n} w_{fk} h_{nk} v_{jfn}^{-1}} \right), \tag{6}$$

$$w_{fk} \leftarrow w_{fk} \left(\frac{\sum_{j,n} h_{nk} q_{jk} p_{jfn} v_{jfn}^{-2}}{\sum_{j,n} h_{nk} q_{jk} v_{jfn}^{-1}} \right), \tag{7}$$

$$h_{nk} \leftarrow h_{nk} \left(\frac{\sum_{j,f} w_{fk} q_{jk} p_{jfn} v_{jfn}^{-2}}{\sum_{j,f} w_{fk} q_{jk} v_{jfn}^{-1}} \right). \tag{8}$$

These MU rules must be repeated several times till some convergence criteria are met.

The NTF model parameters can be initialized and constrained in various ways to enable an accurate representation. In this paper we will define and compare 3 schemes:

1. **Pre-Assigned Estimation:** Given the inactive time-frequency samples, y_{jfn}, $(j, f, n) \in \mathbf{\Omega}$, one can set $h_{nk} = 0$, $\forall (j, f, n) \in \mathbf{\Omega}$ and $k \in \mathcal{K}_j$, as well as $q_{jk} = 0$, $\forall j \in [\![1, J]\!]$ and $k \in \mathcal{K}_j$ where $\mathcal{K}_j$ represents the set of components belonging to

source j and $\sum_{j=1}^{J} \#\mathcal{K}_j = K$. This is the classical method that has been used in various approaches [10–12, 16]. In this case $\mathbf{Q}$ is set such that each component is assigned to only one source and the temporal activation of each component in $\mathbf{H}$ is assigned with respect to the temporal activation of the corresponding source. Hence it is required to assign the components to the sources via the sets $\mathcal{K}_j$ manually. Once the $\mathbf{Q}$ and $\mathbf{H}$ matrices are initialized, the silent temporal samples of the sources are no longer used in the signal estimation step unlike the proposed methods.

2. **Relaxed Estimation:** In this approach a minimum number of components, K_{min}, are assigned to each source through $\mathbf{Q}$ at the initialization and the remainder of the matrix $\mathbf{Q}$ is estimated automatically. When $K_{\text{min}} = 0$, this approach is simply estimating all the model parameters, $\boldsymbol{\theta}$, relying on the likelihood maximization.
3. **Sparse Estimation:** It is often the case that the sources exhibit independent characteristics, hence the representation can be more accurate when the components are assigned only to a single source. The *Relaxed Estimation* does not enforce each component to exclusively contribute to one source. In *Sparse Estimation*, the coefficients in $\mathbf{Q}$ and $\mathbf{H}$ are constrained to be sparse so that a structure of zeros similar to the initialized patterns in the *Pre-Assigned Estimation* can be attained without the need to manually specify how many components are assigned to each source. It should also be noted that unlike *Pre-Assigned* and *Relaxed Estimation*, *Sparse Estimation* cannot be performed by specific initialization, instead the multiplicative update rules for the model parameters must be modified to perform sparse NTF decomposition as described in [23].

2.4. Signal Estimation Criterion

Let $\boldsymbol{\Omega}_{fn} \subset [\![1, J]\!]$ be the set of source indices defined as

$$\boldsymbol{\Omega}_{fn} \triangleq \{j | (j, f, n) \in \boldsymbol{\Omega}\}, \tag{9}$$

and let $\mathbf{s}_{fn} \triangleq [s_{1fn}, \ldots, s_{Jfn}]^T$ and $\#\boldsymbol{\Omega}_{fn}$-length column vector $\mathbf{y}_{fn} \triangleq [y_{jfn}]_{j \in \boldsymbol{\Omega}_{fn}}^T$.

Now we can define an observation vector for each time-frequency point (f, n) as $\mathbf{o}_{fn} \triangleq \left[\mathbf{y}_{fn}^T, x_{fn}\right]^T$. Let $\mathbf{O} = \{\mathbf{o}_{fn}\}_{f,n}$ the set of all observed data. In our approach we estimate the model in the maximum likelihood sense, i.e., maximizing the likelihood of the observed data given the model parameters, which writes:[3]

$$p(\mathbf{O}|\boldsymbol{\theta}) = \prod_{f=1}^{F} \prod_{n=1}^{N} \frac{1}{\left|\pi \boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{o}_{fn}}\right|} \exp\left[-\mathbf{o}_{fn}^H \boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{o}_{fn}}^{-1} \mathbf{o}_{fn}\right], \tag{10}$$

since $\mathbf{o}_{fn}$ may be shown zero-mean complex Gaussian vector with covariance matrix $\boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{o}_{fn}}$ that can be expressed as follows:

$$\boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{o}_{fn}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_{fn}\mathbf{y}_{fn}} & \boldsymbol{\Sigma}_{x_{fn}\mathbf{y}_{fn}}^H \\ \boldsymbol{\Sigma}_{x_{fn}\mathbf{y}_{fn}} & \boldsymbol{\Sigma}_{x_{fn}x_{fn}} \end{bmatrix}, \tag{11}$$

where

$$\boldsymbol{\Sigma}_{\mathbf{y}_{fn}\mathbf{y}_{fn}} = \mathbf{I}^H(\boldsymbol{\Omega}_{fn}) \text{diag}\left(\left[v_{jfn}\right]_j\right) \mathbf{I}(\boldsymbol{\Omega}_{fn}) \tag{12}$$

$$+ \text{diag}\left([\sigma_{b,jfn}^2]_{j \in \boldsymbol{\Omega}_{fn}}\right), \tag{13}$$

$$\boldsymbol{\Sigma}_{x_{fn}\mathbf{y}_{fn}} = [v_{jfn}]_j \mathbf{I}(\boldsymbol{\Omega}_{fn}),\ \boldsymbol{\Sigma}_{x_{fn}x_{fn}} = \sum\nolimits_j v_{jfn}, \tag{14}$$

[3] $\bullet^H$ denotes conjugate transpose.

Algorithm 1 GEM algorithm for Source Separation using NTF model

1: **procedure** SSEPARATION-NTF($\{x_{fn}\}_{f,n}$, $\{y_{jfn}\}_{(j,f,n)\in\boldsymbol{\Omega}}$)
2: Initialize non-negative $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ with respect to *Pre-Assigned*, *Relaxed* or *Sparse Estimation*
3: **repeat**
4: Estimate posterior power spectra, $\widehat{\mathbf{P}} = \mathbb{E}\left\{[|s_{jfn}|^2]_{j,f,n} | \mathbf{O}; \boldsymbol{\theta}\right\}$
5: Update $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ given $\widehat{\mathbf{P}}$
6: **until** convergence criteria met
7: **end procedure**

and $\mathbf{I}(\boldsymbol{\Omega}_{fn})$ is the $J \times \#\boldsymbol{\Omega}_{fn}$ matrix consisting of the columns of $J \times J$ identity matrix $\mathbf{I}_J$ that belong to $\boldsymbol{\Omega}_{fn}$.

2.5. Algorithm Summary

The general flow of the proposed generalized expectation maximization (GEM) algorithm is the same for all three approaches, with the difference of initialization and update of the NTF model parameters as described in Section 2.3. After the initialization, the algorithm iteratively alternates between updating the posterior power spectra, $\widehat{\mathbf{P}}$, given the model parameters and updating the NTF model parameters by the multiplicative update rules given the posterior power spectra, as described in Section 2.3. The posterior power spectra can be computed as

$$\widehat{\mathbf{P}} = \mathbb{E}\left\{[|s_{jfn}|^2]_{j,f,n} | \mathbf{O}; \boldsymbol{\theta}\right\} = |\mathbb{E}\left\{[s_{jfn}]_{j,f,n} | \mathbf{O}; \boldsymbol{\theta}\right\}|^2 + \left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{fn}\mathbf{s}_{fn}}\right]_{f,n} \tag{15}$$

where

$$\mathbb{E}\left\{[s_{jfn}]_{j,f,n} | \mathbf{O}; \boldsymbol{\theta}\right\} = \left[\boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{s}_{fn}}^H \boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{o}_{fn}}^{-1} \mathbf{o}_{fn}\right]_{f,n}, \tag{16}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{fn}\mathbf{s}_{fn}} = \boldsymbol{\Sigma}_{\mathbf{s}_{fn}\mathbf{s}_{fn}} - \boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{s}_{fn}}^H \boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{o}_{fn}}^{-1} \boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{s}_{fn}}, \tag{17}$$

$$\boldsymbol{\Sigma}_{\mathbf{o}_{fn}\mathbf{s}_{fn}} = \begin{bmatrix} v_{1fn} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_{Jfn} \\ v_{1fn} & \cdots & v_{Jfn} \end{bmatrix}, \boldsymbol{\Sigma}_{\mathbf{s}_{fn}\mathbf{s}_{fn}} = \begin{bmatrix} v_{1fn} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_{Jfn} \end{bmatrix}. \tag{18}$$

The overall GEM algorithm steps are summarized in Algorithm 1.

3. EXPERIMENTS

In order to assess the performance of the proposed algorithm, 6 different music mixtures are selected,[4] each composed of 3 sources (bass or drums, guitar and vocals). Source separation is performed on each mixture with the total number of components fixed to 15[5], using 3 different approaches: using equal number of components pre-assigned for each source (*Pre-assigned Estimation*), assigning 2 components per source and estimating the rest of the parameters using the proposed estimation method (*Relaxed Estimation*) and estimating the signals by assuming distribution of components among

[4] The mixtures are taken from the professionally produced music recordings of SiSEC 2015 evaluation campaign (https://sisec.inria.fr/).

[5] This number of components is observed to be optimum for music signals in the experiments performed in [21].

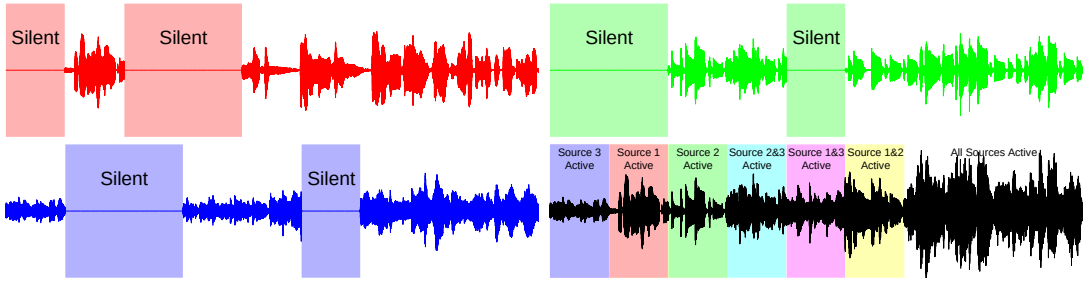

Fig. 1: Source 1 (top left), source 2 (top right), source 3 (bottom left) and the mixture for the mixture (bottom right) 5 in the experiment setup.

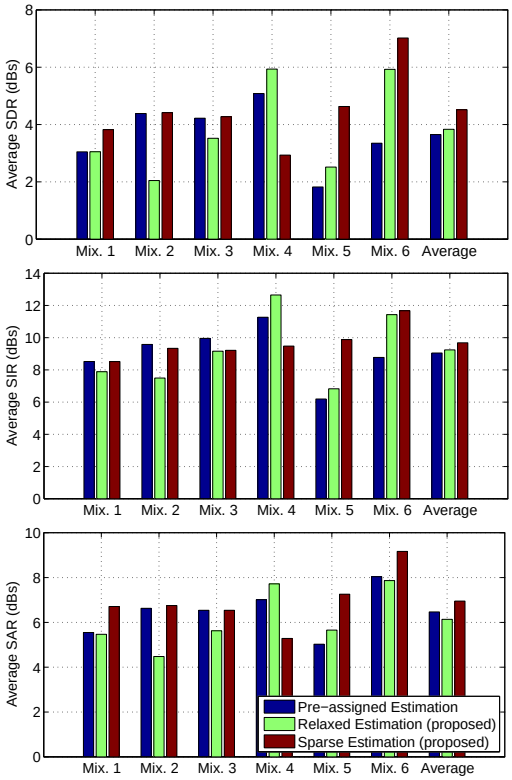

Fig. 2: Comparison of source separation performance for different component allocation methods measured in signal to distortion ratio (SDR, top), source to interference ratio (SIR, middle) and sources to artefacts ratio (SAR, bottom) [24].

the sources are sparse (*Sparse Estimation*). The STFT is computed using a half-overlapping sine window of 1024 samples (64 ms) and the proposed GEM algorithm is run for 500 iterations. The sources in the mixtures are artificially silenced during a percentage of the total time, an example of which is shown in Figure 1.

The overall performance of the compared algorithms can be seen in Figure 2. It can be noticed that the proposed *Sparse Estimation* method often outperforms blindly assigning equal number of components for each source. This can be attributed to the fact that the sources in the test mixtures often have varying complexity, for example the vocals or guitar are known to have greater spectral patterns variability than the bass and drums. Since the proposed method allows uneven distribution of the components among the sources, the performance is often better. The *Relaxed Estimation* method can also be seen to have better or close performance to *Pre-assigned Estimation* for most of the mixtures. This approach also allows automatic assignment of components hence it can benefit from non-uniform distribution of components among the sources. However it does not enforce the components to be exclusively assigned to any of the sources, therefore its performance is usually worse than *Sparse Estimation* which induces exclusivity by enforcing sparsity in the matrix $\mathbf{Q}$.

It should be noted that, even though not displayed in Figure 2, the use of the initialization in *Relaxed Estimation* and the sparsity constraint in *Sparse Estimation* are also tested without the use of silent samples from the sources in the signal estimation step. However the performance was significantly worse than results shown in Figure 2. This demonstrates that the inclusion of the known (silent) samples in the signal estimation is essential for the success of the proposed automatic component allocation algorithms. Another important remark is that, when the known samples of the sources are zero, i.e. the sources are known to be silent, it is necessary to apply the proposed algorithms with non-zero noise variance, $\sigma^2_{b,jfn}$, to avoid numerical errors in matrix inversion during the signal estimation step. Fortunately, this necessity does not create a drawback since the noise variance, $\sigma^2_{b,jfn}$, can be chosen sufficiently small to avoid inaccurate signal reconstructions.

4. CONCLUSION

In this paper, two new methods are presented which enable automatic allocation of NTF components among multiple sources when performing source separation with low-rank NTF models and user guided annotations. The *Relaxed Estimation* method does not impose any structure on the NTF model parameters except assigning a small number of components exclusively to each source and relies on the maximum likelihood estimation using the prior information on the annotations to determine the unknown NTF model. The *Sparse Estimation* method imposes a sparsity prior on the distribution of the components among the sources so that each component is forced to be assigned to a single source. The proposed methods are tested against the classic approach of manually assigning equal number of components to each source on artificially silenced music mixtures, sources of which have variable degrees of complexity. As a result it has been shown that *Sparse Estimation* outperforms the other approaches in most of the cases, and even the *Relaxed Estimation* performs better than the classic approach for many test mixtures.

The proposed algorithms are still not perfect and can perform worse than the classic method in rare cases for which strictly constraining the NTF model is more useful than having larger degrees of freedom in the model.

5. REFERENCES

[1] E. Vincent, S. Araki, F.J. Theis, G. Nolte, P. Bofill, H. 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] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," vol. 20, no. 4, pp. 1118–1133, 2012.

[3] 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.

[4] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected f0 track," in *Latent Variable Analysis and Signal Separation*, Fabian Theis, Andrzej Cichocki, Arie Yeredor, and Michael Zibulevsky, Eds., vol. 7191 of *Lecture Notes in Computer Science*, pp. 438–445. Springer Berlin Heidelberg, 2012.

[5] Z. Rafii, A. Liutkus, and B. Pardo, "A simple user interface system for recovering patterns repeating in time and frequency in mixtures of sounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 271–275.

[6] D. Fitzgerald, "User assisted separation using tensor factorisation," in *20th European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, 2012, pp. 2412–2416.

[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] D. El Badawy, N. Q. K. Duong, and A. Ozerov, "On-the-fly audio source separation," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2014.

[9] Nicholas Bryan, Gautham J. Mysore, and Ge Wang, "Isse: An interactive source separation editor," in *CHI Conference on Human Factors in Computing Systems*, Toronto, Canada, 04/2014 2014, ACM, ACM, .

[10] 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.

[11] B. Wang, "Musical audio stream separation," M.S. thesis, 2009.

[12] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, May 2011, pp. 257–260.

[13] 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.

[14] N. J. Bryan and G. J. Mysore, "An efficient posterior regularized latent variable model for interactive source separation," in *Proc. Int. Conf. on Machine Learning (ICML)*, Atlanta, GA, June 2013.

[15] N.Q.K. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," in *IEEE International Conference on Consumer Electronics (ICCE-Berlin)*, Berlin, Germany, Sept. 2014.

[16] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'14)*, Florence, Italy, May 2014.

[17] A. Lefèvre, F. Glineur, and P.-A. Absil, "A convex formulation for informed source separation in the single channel setting," *Neurocomputing*, vol. 141, pp. 26–36, Oct. 2014.

[18] I.-Y. Jeong and K. Lee, "Informed source separation from monaural music with limited binary time-frequency annotation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[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] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing 13 (NIPS'2000)*, 2001.

[21] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," vol. 18, no. 3, pp. 550–563, Mar. 2010.

[22] Ç. Bilen, A. Ozerov, and P. Pérez, "Compressive sampling-based informed source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2015.

[23] J. Le Roux, F. Weninger, and J. R. Hershey, "Sparse NMF? half-baked or well done?," Tech. Rep., Mitsubishi Electric Research Laboratories, 2015.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," vol. 14, no. 4, pp. 1462–1469, July 2006.