Structured Projective Non Negative Matrix Factorization with drum dictionaries for Harmonic/Percussive Source Separation

Clément Laroche, Matthieu Kowalski, Hélène Papadopoulos, *Member, IEEE,* and Gaël Richard, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a new unconstrained nonnegative matrix factorization method designed to use the multilayer structure of audio signals in order to improve the quality of source separation. Audio signals (and music signals in particular) can be decomposed in two distinct parts: a tonal component which is sparse in frequency and temporally stable and a transient component composed of short term broadband sounds. Our method decomposes the audio signal in sparse orthogonal components that are well suited for representing the tonal component, while the transient part is represented by a regular nonnegative matrix factorization decomposition. Experiments on a large database of real music data show that such decomposition is suitable for audio source separation. Tested in comparison with three state-of-the-art harmonic/percussive decomposition algorithms, the proposed method shows competitive performances.**

***Index Terms*—nonnegative matrix factorization, projective nonnegative matrix factorization, audio source separation, harmonic/percussive decomposition.**

I. Introduction

Non Negative Matrix Factorization (NMF) is a widely used rank reduction method. The goal of NMF is to approximate a data matrix $V \in \mathbb{R}_+^{n\times m}$ as $V \approx \tilde{V} = WH$ with $W \in \mathbb{R}_+^{n\times k}$, $H \in \mathbb{R}_+^{k\times m}$, k being the rank of factorization typically chosen such that $k(n+m) << nm$ [1]. As the data matrix V is usually redundant, the product WH can be thought as a compressed form of V. In audio signal processing, the input data is usually a Time-Frequency (TF) representation such as a short time Fourier transform (STFT) or a constant-Q transform spectrogram. The matrix W is a *dictionary* or a set of *patterns* that codes the frequency information of the data. H is the *activation matrix* and contains the *expansion coefficients* that code the temporal information. This decomposition technique has been applied with great success in various audio signal processing tasks such as automatic transcription [2], [3], audio source separation [4], [5], multi-pitch estimation [6] and music instruments recognition [7].

Numerous schemes have been proposed in the NMF framework in the context of audio source separation. In the case where the signals are composed of many sources, the plain NMF does not give satisfying results and it is often necessary to rely, for instance, on supervised algorithms that exploit training data or prior information in order to guide the decomposition process. Also, information from the scores or from midi signals [2] can be used to initialize the learning process. The downside of this approach is that it requires a well organized prior information that is not always available. Another approach consists in performing prior training on specific databases. A dictionary matrix W_{train} can be learned from a large database in order to separate an instrument [8], [9]. A common method to build a dictionary for NMF is to perform a decomposition on a large training set. After the convergence, the W matrix from the decomposition is used as the dictionary matrix W_{train} in the separation [8]. Another method is detailed in [9], where a dictionary matrix is created by extracting template spectra from isolated drum samples. The dictionary is then used in a NMF decomposition to perform drum transcription. This method requires minimum tuning from the user. However, the dictionary should match the target instrument for satisfying performances.

On the other hand, unsupervised algorithms rely on parametric physical models of the instruments and they count on specific constraints deduced from the characteristics of the processed signals in order to perform the decomposition. For example, harmonic instruments tend to have regular activations and are slowly varying over time so enforcing temporal smoothness improves the physical meaning of the decomposition [10]. Similarly in [11], Canadas & al. use four constraints to achieve a specific harmonic/percussive decomposition. The main drawback of parametrized methods is that the hyper parameters are difficult to tune, especially if the model is composed of numerous parameters.

Concurrently, other unsupervised methods aim at underlining some mathematical properties of the decomposition, like the orthogonality between the nonnegative basis functions (or patterns). The Projective NMF (PNMF) and the Orthogonal NMF (ONMF) are typical examples of such approaches. The PNMF has been used with success in image processing [12] for feature extraction and clustering [13]. It reveals interesting properties in practice: a higher efficiency for clustering than the NMF [12] as well as the generation of a much sparser decomposition [13]. These inherent properties are particularly interesting for audio source separation as shown in [11], [14]. The main advantage of these approaches compared to other unsupervised methods is that sparsity or orthogonality are

C. Laroche and G. Richard are with Institut Mines-Télécom, Télécom ParisTech, CNRS-LTCI, 37-39 rue Dareau, 75014 Paris, France.

C. Laroche, M. Kowalski and H. Papadopoulos are with Laboratoire des Signaux et Systèmes, UMR 8506 CNRS - CENTRALESUPELEC - Univ Paris-Sud, 91192 Gif-sur-Yvette Cedex, France.

This work was supported by a grant from DIGITEO.

obtained as intrinsic properties, so they avoid a tedious and often unsatisfactory hyperparameter tuning stage. However, these approaches do not have a sufficient flexibility to properly represent the complexity of an audio scene composed of multiple and concurrent harmonic and percussive sources.

In this article, we propose a semi-supervised decomposition technique suitable for harmonic/percussive source separation of non vocal monophonic instrumental signals. The method takes advantage of the sparse decomposition of the PNMF but allows a better representation of complex audio signals. More precisely, the initial nearly-orthogonal decomposition obtained by the PNMF is extended by non-orthogonal components that prove to be particularly relevant to represent percussive or transient signals. The sparse and orthogonal components of the PNMF are prone to extract the harmonic instruments well localized in frequency, while the percussive part with flat spectra are extracted by the NMF components. However, as shown in [15] this direct approach obtains limited performances in a blind source separation task. To force the non-orthogonal components to extract the drum signal, we use a dictionary specific to drum signals using the same technique as in [9]. The contributions of this article are threefold

1) The Structured Projected Nonnegative Matrix Factorization (SPNMF) method is introduced, as well as a simple multiplicative algorithm to solve the SPNMF problem.
2) We offer a comparison between two different training methods to establish the best drum dictionary.
3) We perform a benchmark against three state of the art harmonic/percussive methods on a large database.

The paper is organized as follows. In Section II, we describe the problem of harmonic/percussive source separation and we explain in details some recent state-of-the-art methods. The proposed SPNMF is then introduced in Section III. We present our experimental protocol and the results obtained on synthetic and real audio signals in Section IV. Section V is a benchmark with state of the art methods. Finally, some conclusions are drawn in Section VI.

II. Context of harmonic/percussive decomposition

Harmonic/percussive source separation is based on the observation that harmonic instruments are tonal sounds very localized in frequency that can be modeled by a sum of sinusoidal waves and percussive instruments are wide band transient signals.

Harmonic instruments can be classified into several categories such as for example struck string instruments (piano, harpsichord, ...), plucked string instruments (guitar, mandolin, ...) or brass instruments (trumpet, tuba, ...) [16]. These instruments can be modeled as a sum of three components: the Sinusoidal part, the Transients and the Residual, so called Sines + Transients + Noise (STN) models [17]. The transient of the harmonic instruments show some percussive properties (i.e., fast attack and fast decay). Taking into account the transient part of harmonic instruments is a challenging task that will not be treated this article. Here, as in other article of the literature, we rely on the hypothesis that most of the energy of the harmonic signal is in the tonal part.

Harmonic/percussive source separation has numerous applications as a preprocessing step. For example most multi-pitch estimation models [18], instruments recognition [19] and melody extraction [20] algorithms are much more efficient if the influence of the percussive sources is diminished. Indeed, these algorithms often rely on the analysis of the harmonic structures that are blurred by the percussive instruments. Similarly, beat tracking [21] and drum transcription algorithms [22] are more accurate if the harmonic instruments are not part of the signal. Finally, using the Harmonic/Percussive Source Separation (HPSS) algorithm [23] as a preprocessing step increases the performance for singing pitch extraction and voice separation [24]. This has motivated the development of methods for harmonic/percussive sound separation in the music signal processing area. Numerous other methods exist but in this work, we will focus on three methods that are particularly relevant to our benchmark.

A well known unsupervised method to extract the harmonic/percussive components consists in applying a median filtering on the spectrogram of the audio signal [23], [25]. The filtering is made across the temporal atoms to diminish the transient sounds in order to extract the harmonic components. Mutually, the filtering is made across frequency to reduce the harmonic tonal sounds and to enhance the percussive instruments. The assumption made is that the harmonics are considered to be outliers in a temporal frame that contains a mixture of percussive and pitched instruments. Similarly, the percussive onsets are considered to be outliers in a frequency frame. This method is often used in the *Music Information Retrieval* community as it does not require any parameter tuning and is computationally effective. However, on a small scale test conducted by [11], this simple approach does not give the best separation results.

Another unsupervised method uses a NMF decomposition with specific constraints to distinguish the harmonic part from the percussive components [11]. A simple approximation is that percussive instruments are transient sounds with regular spectra (wide band signals) whereas harmonic instruments are tonal sounds with harmonic sparse spectra. From this observation, a frequency regularity and a temporal sparsity constraints are applied during the optimization process to extract the percussive instruments and vice-versa a temporal regularity and a frequency sparsity constraints are applied to extract the harmonic instruments. This method gives good results compared to other state-of-the-art methods but the hyper-parameters tuning is a tedious process and results depend heavily on the training database.

Finally in [26], a drum source separation is done using a Non-Negative Matrix Partial Co-Factorization (NMPCF). The spectrogram of the signal and the drum-only data (obtained from prior learning) are simultaneously decomposed in order to determine common basis vectors that capture the spectral and temporal characteristics of drum sources. The shared dictionary matrix retrieves the drum signal. However it must be chosen carefully in order to obtain satisfying results. The percussive part of the decomposition is constrained while the harmonic part is completely unconstrained. As a result, it tends to decompose a lot of information (i.e., the harmonic

part contains some percussive instruments) from the signal and the decomposition is not satisfactory. In the NMPCF, the dictionary is not fixed. We will compare this method to the SPNMF were the dictionary is fixed.

III. Structured projective NMF (SPNMF)

In this section we present our semi-supervised algorithm for harmonic/percussive source separation.

A. Presentation of the orthogonal and projective NMF

Using a squared Euclidean distance between the data matrix V and its approximation WH, the NMF problem reads:

$$\min_{W,H\geq 0} \|V - WH\|^2 \ ,$$

where $\|.\|^2$ is the squared Euclidean distance.

The aim of the PNMF is to find a non negative projection matrix $P \in \mathbb{R}_+^{n\times n}$ such that $V \approx \tilde{V} = PV$. In [27] Yuan & al. propose to seek P as an approximative projection matrix under the form $P = WW^T$ with $W \in \mathbb{R}_+^{n\times k}$ with $k \leqslant n$. The PNMF problem reads :

$$\min_{W\geqslant 0} ||V - WW^TV||^2 \tag{1}$$

PNMF is similar to the NMF problem and can be simply obtained by replacing the activation matrix H by W^TV. It is shown in [13] that the PNMF gives a much sparser decomposition than the NMF.

Another very similar approach is the ONMF [12]. It consists in solving the following problem:

$$\min_{W\geqslant 0, H\geqslant 0} ||V - WH||^2 \quad \text{s.t.} \quad W^TW = I_k \tag{2}$$

In this method, orthogonality between nonnegative basis functions is enforced during the optimization process. In theory, it seems that the PNMF and the ONMF lead to similar decompositions, as the W matrix estimated by the PNMF is almost orthogonal (i.e., $\|W^TW - I_k\|^2$ is small). However in practice, enforcing the orthogonality between the base at every iteration is a constraint too strong to decompose audio signal [15].

The sparsity of the dictionary matrix is an interesting property for the decomposition of audio signals and especially for the decomposition of harmonic instruments with very localized harmonic spectra. Contrary to the NMF, the sparsity of the PNMF is an inherent features of the decomposition. These key properties of the PNMF motivated us to decompose the harmonic instruments using orthogonal basis functions.

B. Principle of the SPNMF

The orthogonal basis functions of the PNMF are not flexible enough to decompose a complex audio signal. As stated in [11], harmonic instruments have sparse basis functions whereas percussive instruments have much flatter spectra. As the columns of W are orthogonal, when two sources overlap in the Time-Frequency (TF) plane only one basis function will represent the mixture which is not adequate for efficient separation. To overcome this problem, we propose to add a standard NMF decomposition term to the PNMF. We can expect that most of the harmonic components will be represented by the orthogonal part while the percussive ones will be the regular NMF components. Using a similar model to preliminary work [15], let V be the magnitude spectrogram of the input data. The model is then given by

$$V \approx \tilde{V} = V_H + V_P, \tag{3}$$

with V_P the spectrogram of the percussive part and V_H the spectrogram of the harmonic part. V_H is approximated by the PNMF decomposition while W_P is decomposed by some NMF components as :

$$V \approx \tilde{V} = W_H W_H^T V + W_P H_P. \tag{4}$$

The data matrix is approximated by an almost orthogonal sparse part that codes the harmonic instruments $V_H = W_H W_H^T V$ and a non constrained NMF part that codes the percussive instruments $V_P = W_P H_P$. The main advantage of the SPNMF results is in the fact that the method has few parameters compared to constrained NMF [11]and NMPCF [26] while still obtaining similar results [15].

C. Algorithm Optimization

In order to achieve such a decomposition, we can use a measure of fit $D(x|y)$ between the data matrix V and the estimated matrix $\tilde{V}$. $D(x|y)$ is a scalar cost function and in this article we compare in Section IV-D the use of the Euclidean distance (Euc), the Kullback Leiber (KL) divergence and the Itakura Saito (IS) divergence which are three commonly used divergences in the NMF framework.

The SPNMF model gives the following cost function :

$$\min_{W_H, W_P, H_P\geq 0} D(V|W_H W_H^T V + W_P H_P) \tag{5}$$

A solution of this problem can be obtained by iterative multiplicative update rules following the same strategy as in [27], [28] which consists in splitting the gradient with respect to (wrt) one variable (here W_H for exemple) $\nabla_{W_H} D(V|\tilde{V})$ in its positive $[\nabla_{W_H} D(V|\tilde{V})]^+$ and negative parts $[\nabla_{W_H} D(V|\tilde{V})]^-$. The multiplicative updates for SPNMF are then given by:

$$W_H \leftarrow W_H \otimes \frac{[\nabla_{W_H} D(V|\tilde{V})]^-}{[\nabla_{W_H} D(V|\tilde{V})]^+},$$

where $\otimes$ is the Hadamard product or element-wise product. Details of the equations for the Euc distance, KL and IS divergences are given in the annex in Sections VI-A, VI-B and VI-C respectively. The optimization of the SPNMF algorithm is done sequentially, as described in Algorithm 1.

D. SPNMF with fixed dictionary

A fully unsupervised SPNMF model does not allow for a satisfying harmonic/percussive source separation [15]. To alleviate this problem, we use here a fixed drum dictionary W_p for the percussive part of the SPNMF which is created using the drum database ENST-Drums [29]. In this database three professional drum players specialized in a specific type

Input: $V \in \mathbb{R}_+^{m\times n}$ Output: $W \in \mathbb{R}_+^{m\times k}$, $W_P \in \mathbb{R}_+^{m\times e}$ and $H_P \in \mathbb{R}_+^{e\times n}$ Initialization;
while $i \leq$ *number of iterations* **do**
$W_P \leftarrow W_P \otimes \frac{[\nabla_{W_P} D(V|\tilde{V})]^-}{[\nabla_{W_P} D(V|\tilde{V})]^+}$
$H_P \leftarrow H_P \otimes \frac{[\nabla_{H_P} D(V|\tilde{V})]^-}{[\nabla_{H_P} D(V|\tilde{V})]^+}$
$W_H \leftarrow W_H \otimes \frac{[\nabla_{W_H} D(V|\tilde{V})]^-}{[\nabla_{W_H} D(V|\tilde{V})]^+}$
$i = i + 1$
end
$V_P = W_P H_P$ and $V_H = W_H W_H^T V$

Algorithm 1: SPNMF algorithm with multiplicative update rules.

of music have been recorded. Each drummer uses a specific drum set from a small one (two toms and two cymbals) to a full rock drum kit (four toms and five cymbals).

Two different ways can be used to build a dictionary. The first one is to perform a NMF decomposition on a large database [8]. The recording of the three drummers are concatenated independently to obtain three long audio drum signals of each drum kit. We execute a NMF on the spectrogram of the signals. The rank of the decomposition is chosen as $k = 12$ for each of the signals. We consider that each element of the drum kit is represented by two basis functions. This method is effective at giving a dictionary specific to an instrument and the size of the dictionary is maintained low so it does not increase the computation time. However, the template of the dictionary does not represent a single element of the drum kit so it is not possible to perform direct drum transcription.

A second way to build the dictionary is to directly use the STFT of a drum signal [9]. The ENST-Drum database contains audio files where elements of the drum kit are played independently and that we used to build another types of drum dictionary. This method allows having a specific dictionary, yet the matrix is redundant and very large which increases the computation time. Both learning methods are compared in Section IV-E.

The main advantage of the proposed method over the previously mentioned concurrent approaches is that it can take into account some of the percussive instruments that have a sparse spectrum. The bass drum and the toms have almost harmonic spectra with most energy in the low frequency range. These sounds are consequently hard to extract while enforcing frequency regularity of the percussive part like in [11], [25].

The SPNMF algorithm with the fixed dictionary matrix is described by Algorithm 2.

E. Signal reconstruction

The percussive signal $x_p(t)$ is synthesized using the magnitude percussive spectrogram $V_P = W_P H_P$. To reconstruct the phase of the percussive part, we use a generalized Wiener filter [30] to create a percussive mask as:

$$\mathcal{M}_P = \frac{V_P^\alpha}{V_M^\alpha + V_P^\alpha}, \tag{6}$$

Input: $V \in \mathbb{R}_+^{m\times n}$ Output: $W \in \mathbb{R}_+^{m\times k}$, $W_{train} \in \mathbb{R}_+^{m\times e}$ and $H \in \mathbb{R}_+^{e\times n}$ Initialization;
while $i \leq$ *number of iterations* **do**
$H_P \leftarrow H_P \otimes \frac{[\nabla_{H_P} D(V|\tilde{V})]^-}{[\nabla_{H_P} D(V|\tilde{V})]^+}$
$W_H \leftarrow W_H \otimes \frac{[\nabla_{W_H} D(V|\tilde{V})]^-}{[\nabla_{W_H} D(V|\tilde{V})]^+}$
$i = i + 1$
end
$V_P = W_{train} H_P$ and $V_H = W_H W_H^T V$

Algorithm 2: SPNMF with the drum dictionary matrix.

, where $\alpha \in (1, 2)$. To retrieve the percussive signal as

$$x_p(t) = InverseSTFT(\mathcal{M}_P \otimes X), \tag{7}$$

where X is the complex spectrogram of the mixture.

Similarly for the harmonic part, we obtain:

$$\mathcal{M}_H = \frac{V_H^\alpha}{V_M^\alpha + V_P^\alpha}, \tag{8}$$

and:

$$x_h(t) = InverseSTFT(\mathcal{M}_H \otimes X). \tag{9}$$

IV. Experimental Validation of SPNMF

In this section we conduct a set of experiments to assess the merits of the proposed method. We first perform a test on a synthetic signal to validate the model of SPNMF in section IV-A. We then set-up the SPNMF to perform efficient source separation by quantifying the influence of the rank of factorization in section IV-C, the effect of different divergences in section IV-D and finally the performance of the separation with different types of dictionaries in section IV-E.

A. Synthetic Tests

To illustrate how the straight SPNMF (Algorithme 1) works, we use a simple synthetic signal. The test signal models a mix of harmonic and percussive instruments. The harmonic part is simulated by a sum of sine waves that overlap in time and frequency. The first signal simulates a $C(3)$ with fundamental frequency $f_0 = 131$ Hz, the other one a $B(4)$ with $f_0 = 492$ Hz. To simulate the percussive part, we add 0.1 s of Gaussian white noise for the first two second. For the last two seconds, we add 0.3 s of Gaussian white noise filtered by a high-pass filter. The signal is 5 s long and the sampling rate is 4000 Hz. We compute the Short Time Fourier Transform (STFT) with a 512 sample-long (0.128 s) Hann analysis window and a 50% overlap. The spectrogram of the signal is represented in Figure 1. As our input signal has four sources, we expect that one source can be represented by one component and therefore, a model of rank 4 ($k = 4$) should adequately model the signals. More precisely, for the NMF and the PNMF we chose $k = 4$ and for the SPNMF $k' = 2$ and $e = 2$. The choice of the rank of factorization is an important variable of the problem. In this case, we select it in order to illustrate the performance of the method. We will further discuss the importance of the choice of the rank of factorization in

Section IV-C. We compare the SPNMF with the PNMF and the NMF using the KL distance with multiplicative update rules as stated in [31]. The three algorithms are initialized with the same random positive matrices $W_{ini} \in \mathbb{R}^{n\times k}$ and $H_{ini} \in \mathbb{R}_{+}^{k\times m}$.

The results of the decomposition are presented in Figure 2. The dictionary and activation matrices show the separation performance of the three methods. The NMF does not separate correctly the four components. By looking at the columns 2 and 3 of the dictionary matrix W on Figure 2, the filtered Gaussian white noise and the $C(3)$ are separated in the same components which does not correspond to the expected result. For the PNMF, the orthogonal components do not succeed to represent the two noises correctly. They are extracted in the same component and the total reconstruction error of the PNMF is high. In this example, the SPNMF extracts the four components with the highest accuracy and performs better than the other methods. The two harmonic components are extracted in the orthogonal part (i.e., the columns 1 and 2 of the dictionary matrix) while the percussive components are extracted by the NMF part (columns 3 and 4). The SPNMF outperforms the two other methods and shows therefore the potential of the proposed algorithm for harmonic/percussive source separation.

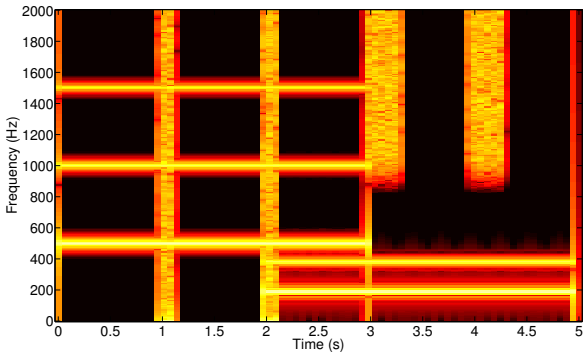

Fig. 1: Spectrogram of the synthetic test signal.

B. Protocol and details of the test database

To study the impact of using learned dictionaries in SPNMF (Algorithm 2), we run several tests on the public SiSec database from [32]. This database is composed of polyphonic real-world music excerpts. Each music signal contains percussive, harmonic instruments and vocals. It consists of four recordings whose durations range from 14 to 24 s. Our goal is to perform a harmonic/percussive decomposition. Thus, following [11], we do not consider the vocal part and we build mixture signals only from the percussive and harmonic instruments. All the signals are sampled at $44.1kHz$. We compute the STFT with a 1024 and 2048 sample-long Hann window with a 50% overlap. Three tests are run on these data:

1) The first test aims at assessing the robustness of the SPNMF with respect to the rank of the PNMF part.
2) The second test is to evaluate which of the three divergences (Euc, KL and IS respectively) give the best harmonic/percussive decomposition results.
3) The last test shows the influence of the dictionary on the separation performance.

Note that the database we use for tuning the proposed method is different from the one in the evaluation phase in order to prevent any possible over-training and therefore to get the most accurate and fair comparison. In order to evaluate and compare the results we then compute the common Signal to Distortion Ratio/Signal to Interference Ration/Signal to Artefact Ratio (SDR/SIR/SAR) metrics for blind source separation with the BSS-Eval toolbox [33].

C. Robustness wrt the rank of the harmonic part

In the case where we use a fixed dictionary matrix, the only parameter of the algorithm is the rank of factorization of the harmonic part. In this experiment, we use the SP-NMF algorithm with the fixed dictionary obtained from the STFT of a drum signal as described in Section III-D. The algorithms are implemented using the multiplicative update rules from VI-A, VI-B and VI-C and they all are initialized with the same random non-negative matrices. We display the mean value of the separation results on Figure 3. When the rank of factorization is small, the Euclidean distance and the KL divergence do not give satisfying results. However, for $r >= 100$, the results for both distances remain stable. With the IS divergence, the results seem independent of the rank of the factorization.

The optimization process of SPNMF is straightforward thanks to the robustness of the method wrt the rank of factorization. The number of components that can be decomposed by orthogonal basis functions is limited, and increasing the rank of factorization does not perturb the results as the harmonic part has to be orthogonal. For the rest of the article, the rank of factorization will be set to $r = 100$ for all methods.

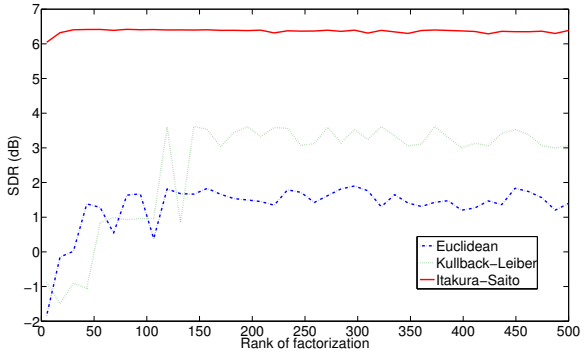

Fig. 3: Optimization of the rank of factorization with the three divergences.

D. Influence of the divergence

In this section we discuss the influence of the divergence in the results of the SPNMF algorithm. It has been established that the IS divergence is well suited for audio signal

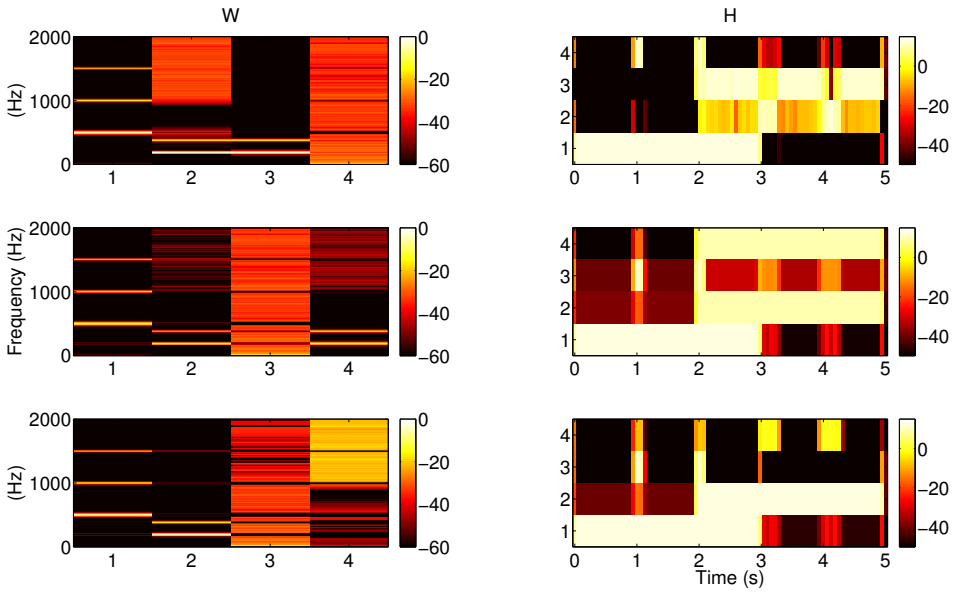

Fig. 2: Results of the decomposition of the NMF (top matrices), PNMF (middle matrices) and SPNMF (bottom matrices).

decomposition [34], even if it does not always lead to superior separation performance [11]. In this section, we perform a comparison of the three divergences on the SiSec database. Also, we compare two different window lengths (1024 samples and 2048 samples) for the Fourier transform as it showed interesting results. We display on Figures 4 and 5 the mean of the results of the three algorithms computed on the SiSec database. Each box-plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles while the whiskers indicate the minimum and maximum values.

When the analysis window length is small, the percussive instruments are well represented and the energy is localized. Using a longer window spreads the percussive energy while the tonal components are well separated in the TF domain.

When the window size is small, Figure 4 shows that the percussive decomposition is better for the Euclidean distance and the KL divergences. However, the harmonic components are not well separated in the TF domain and the orthogonal part of SPNMF does not perform a good separation. In the case of a long window, Figure 5 shows that the IS divergence works better than the other divergences. The orthogonal part is more effective to extract the harmonic components as the finer frequency resolution allows for a better separation in the TF domain. The IS divergence is scale invariant. It means that the low energy components of the spectrogram bear the same relative importance as the higher ones which allows a good extraction of the percussive instruments even if the energy is spread temporally.

For the rest of the article, we will use the SPNMF algorithm with the IS divergence and a window size of 2048 samples for the STFT.

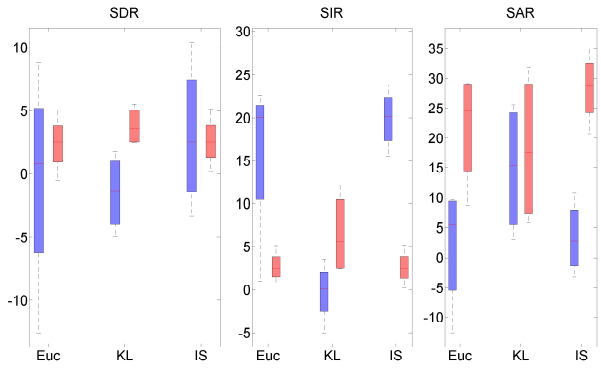

Fig. 4: SDR, SIR and SAR of harmonic (left bar)/percussive (right bar) estimated sources on the SiSec database with a window frame of 1024 samples.

E. Influence of the dictionary

We now discuss the influence of the dictionary. We compare the two methods described in Section III-D. In addition to that, we test a third dictionary is made by the concatenation of the first two dictionaries. The more information contained in the dictionary, the likely the decomposition to properly extract the percussive part. On some signals, the algorithm is not able to extract a lot of energy from the mixture as no atom from the dictionary correspond to any of the percussive signal. We display on Figure 6 the SDR results of the decomposition using the NMF dictionary, the STFT dictionary as well as the concatenated dictionary. The SAR and SIR are postponed in Appendix VI-D on Figures 11 and 12 respectively. In our tests, the results with the concatenated dictionary give the highest score. Indeed, the concatenated dictionary contains the largest amount of information and as a result, obtains the

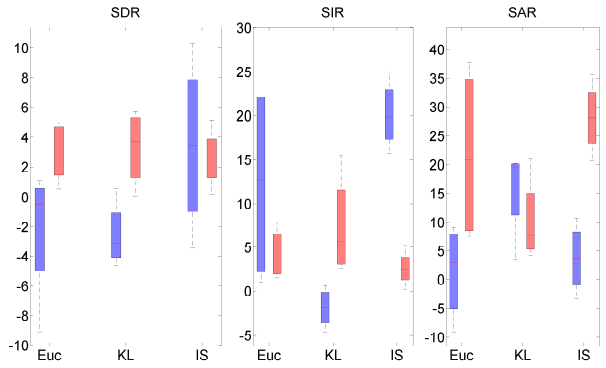

Fig. 5: SDR, SIR and SAR of harmonic (left bar)/percussive (right bar) estimated sources on the SiSec database with a window frame of 2048 samples.

best separation. As the dictionary is fixed, it is important to have a large dictionary to be able to extract a large type of percussive instruments. The STFT and the NMF dictionaries give results similar to each other. The two dictionaries contain complementary information that allow for a better separation while they are concatenated. In the tests we will conduct later in Section V on a large database, we will use the concatenated dictionary as it contains the largest amount of information.

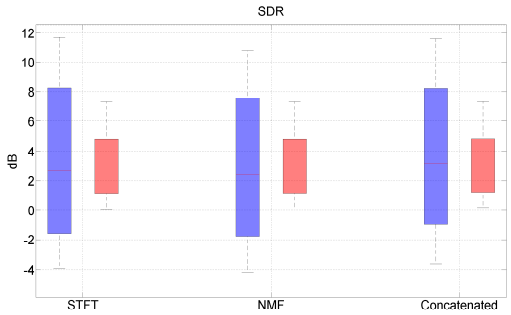

Fig. 6: SDR of harmonic (left bar)/percussive (right bar) estimated sources on the SiSec database with a STFT dictionary, a NMF dictionary and the concatenation of the two.

V. State of the art Benchmark

In this section, we compare the proposed method with three state of the art methods on a large evaluation database. We first detail the database used for evaluation and describe the state of the art methods used for comparison in Section V-A and V-B respectively.

A. Database

For evaluation, we use the Medley-dB [35] database that is composed of polyphonic real-world music excerpts. It has 122 music signals and 89 of them contain percussive instruments, harmonic instruments and vocals. In our tests, the signals that do not contain a percussive part are excluded from evaluation. Because our goal is to perform a harmonic/percussive decomposition, the vocal part is omitted, as in [11]. All the signals are sampled at $44.1kHz$.

B. State of the art methods

We compare here the SPNMF with the fixed dictionary matrix to three other recent state of the art methods: constrained NMF (CoNMF) [11], HPSS [23] and NMPCF [26]. Constrained NMF and NMPCF are re-implemented in this paper and the HPSS implementation is taken from [36].

HPSS is a state of the art, versatile and computationally efficient method. It is widely used in the Music Information Retrieval community and is a good baseline for comparison. The constrained NMF algorithm is the most recent method for harmonic/percussive separation. It gives good results on a small scale test, however the robustness of the algorithm has not been tested yet in a large scale experiment. Finally the NMPCF, similarly to our method, uses a drum dictionary to guide the percussive estimation but the harmonic part is totally unconstrained.

C. Results

Figures 7, 8 and 9 show the SDR, SIR and SAR results of the four methods on the selected 89 songs of the original Medley-dB database [35]. The results on the entire database show that all four methods extract the harmonic instruments much better than the percussive instruments. Our explanation is that all methods rely on Wiener filtering for phase reconstruction (see Equation (8)). As the percussive instruments have flat spectra, the percussive mask is a non sparse matrix and small estimation errors drastically decrease the results of the percussive instruments. This tendency is not visible on small scale tests (see Figure 6).

Figure 7 shows that SPNMF obtains on average the highest separation score for the percussive, harmonic and mean SDR. However, the variance of the results of SPNMF is higher than for the other algorithms. Some songs of the database contain percussive instruments that are not present in the learning database ENST-Drums, such as the tambourine, the bongo, the gong and electronic drums. Because the dictionary is fixed, these percussive instruments are not correctly decomposed by the SPNMF. Some songs are well separated while other obtain much lower results since the percussive part is not well decomposed. This induces an increase of the variance of the results.

The NMPCF, also based on trained data, is more robust than the SPNMF because the dictionary that extract the drums is not fixed. It allows more freedom and the results are more consistent even if some percussive instruments are not in the learning database. However, the mean score is lower than in the case of the SPNMF.

The HPSS results obtained in our tests are unsatisfying. A wide variety of harmonic instruments in the database have really strong transients and rich harmonic spectra (distorted electric guitar, glockenspiel...). Similarly, some percussive instruments have sparse basis functions localized in the low frequency range (bass drum, bongo, toms...). Because of that,

HPSS fails to extract these instruments in the appropriate harmonic/percussive parts. On average, it is able to correctly separate the percussive part (with relatively high SDR and the highest SIR), but it shows a very low SAR compared to the other methods. Similar outcomes have been observed in [11] for HPSS.

The constrained NMF algorithm relies on the same hypothesis than HPSS and the results are lower than those of SPNMF. Some transients of the harmonic instruments are decomposed in the percussive part, and some percussive instruments (mainly in the low frequency range) are decomposed in the harmonic part. The method is still competitive in the large scale test.

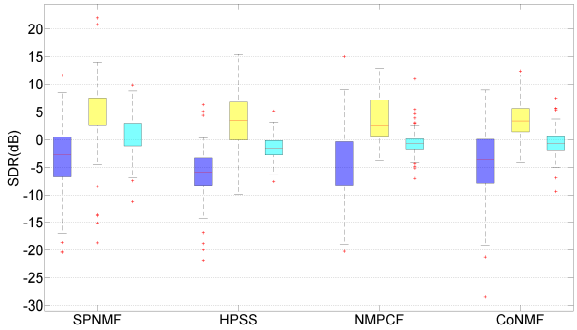

Fig. 7: SDR for percussive/left, harmonic/middle, mean/right separation results on the database for the four methods.

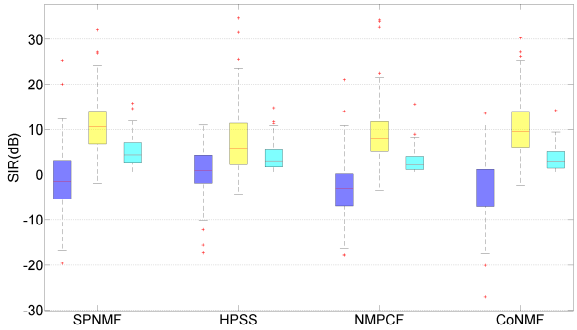

Fig. 8: SIR for percussive/left, harmonic/middle, mean/right separation results on the database for the four methods.

D. Results on a genre specific database

The individual results on most of the songs of the database are similar to the average results. However, some interesting results were found on specific genres of music. Here we present the results obtained on the 14 songs of the "Electronic/Fusion" music genre. These songs for the most part have a lot of silence and some solo parts played by only one instrument. Also, on some songs, the electronic drum repeats the same pattern during the whole song resulting in a very redundant drum part. The SDR results on the sub-database are displayed on Figure 10 (the SIR and SAR results

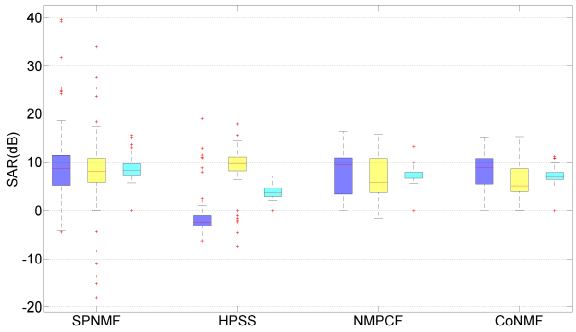

Fig. 9: SAR for percussive/left, harmonic/middle, mean/right separation results on the database for the four methods.

are represented in Appendix VI-E on Figures 13 and 14 respectively).

The HPSS method gives competitive results, with a low variance for the percussive results and a good overall mean. The HPSS obtains consistent results throughout the database. The results on the genre specific database are significantly better than the ones on the whole database. It reflects the fact that the harmonic/percussive instruments are easier to separate on this genre of songs.

The results of the NMPCF are the lowest of the four methods. The unconstrained harmonic part gives the NMPCF a higher degree of freedom which decreases the score as the information is unequally distributed in the harmonic and percussive layers depending on the signal to be decomposed.

Finally the constrained NMF does not obtain satisfying results on this sub-database either. The hyper-parameters are set to the optimal values obtained on a training database of another genre. Because of that, the value of the parameters are not set correctly and similarly to the NMPCF, the information is not distributed in the appropriate harmonic/percussive parts.

On this sub-database, the SPNMF clearly outperforms the other methods. Similarly to Section V-C, the percussive decomposition of the SPNMF has high variance because some of the instruments are not in the learning database. However, the mean of the percussive decomposition is significantly higher than the constrained NMF and the NMPCF. Furthermore, the harmonic decomposition and the mean results of the SPNMF are clearly above all the other methods. The SPNMF is effective to extract the redundant drum parts. Likewise, as the drum dictionary is fixed, it is unlikely for the percussive part to extract harmonic components. As the columns of W_H are orthogonal, it is also unlikely for the harmonic part to extract percussive components. Contrary to the other algorithms, when the harmonic or percussive instruments are playing alone, the SPNMF does not extract any information in the percussive nor the harmonic part.

E. Discussion

The results on the entire database give us insightful information. The HPSS and the constrained NMF rely on the hypothesis that harmonic instruments have sparse tonal spectrogram

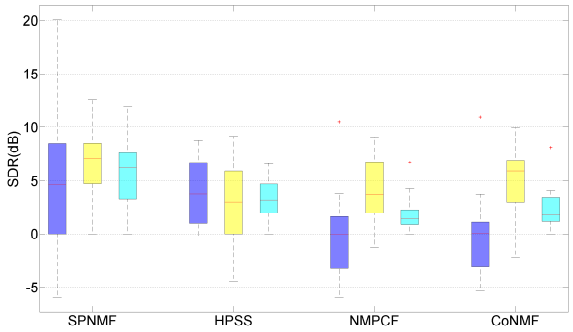

Fig. 10: SDR for percussive/left, harmonic/middle, mean/right separation results on the Electronic/Fusion songs for the four methods.

and that percussive instruments have flat transient spectra. They utilize two different methods to extract the instruments the complementary median filtering, and constraints on the NMF decomposition. The NMPCF uses prior learning to extract the percussive instruments in a specific part while the harmonic instruments are decomposed in an unconstrained layer. The SPNMF combines both techniques from the previous state of the art methods. It uses thus prior learning to extract the percussive instruments while the harmonic parts are extracted by the sparse PNMF components.

Each of the tested methods has its own advantages and drawbacks. The HPSS is the easiest and the fastest method to implement and it does not require any hyper-parameter tuning. The results of the HPSS can be competitive when the harmonic instruments have smooth transients (i.e., sustained instruments such as the flute, the violin) and the percussive instruments have flat spectra (i.e., cymbal, snare drum). However, when the harmonic instruments have strong transients (glockenspiel, piano) and the percussive instruments have sparse spectra (bass drum, bongo) the HPSS does not give good results.

The constrained NMF is based on the same hypothesis than the HPSS and has the same issue. Fine tuning of the hyper-parameters can alleviate the problem mentioned above but it is a tedious process and is not possible in the case of blind source separation. Our tests on a large database show that the constrained NMF is not robust enough for a wide variability of the analyzed signals.

Contrary to the results obtained in [11] the NMPCF algorithm gives competitive results compare to the HPSS and the constrained NMF. However, as it uses training to guide the decomposition process, it requires a wide variety of information to perform on a large scale test. If the training database cannot contain sufficient information, the results will not be satisfying.

On the large scale test, the SPNMF outperforms the other methods. It is able to extract the harmonic and the percussive instruments with higher score for the SDR, the SAR and the SIR. Using prior dictionary learning with a physical model on the harmonic instruments help to separate sources with much better accuracy.

In our test, the training database of the SPNMF and the NMPCF is only composed of drums sounds. The database [35] contains a wide variety of percussive instruments that are not in the training database. However, we decided not to include these types of percussion in the training database as we wanted to have a comparable computation time between the four methods and to test the robustness of the supervised methods when a percussive signal is not in the database.

VI. Conclusion

In this article, we demonstrate that SPNMF is a very promising model for harmonic/percussive decomposition. Indeed, the SPNMF outperforms three other state of the methods on the medley-dB database [35]. Carrying out an evaluation on a large database allowed us to compare more accurately the performance of the four methods on a large variety of music signals.

We can say that the information from the drum dictionary built from the database ENST-Drums [29] is not sufficient to perform a harmonic/percussive source separation on a large scale. Depending of the style of music, some drums share similarities. A possible improvement would be to build genre specific drum dictionaries. In this way, the computation time would be reasonable as the amount information could be reduced and the templates of the dictionary could be a lot more focused on specific type of drums.

Annexe

A. Euclidean distance

The euclidean distance gives us the problem,

$$\min_{W_1,W_2,H_2\geq 0} \|V - W_1W_1^TV + W_2H_2\|^2.$$

The gradient wrt W_1 gives the update

$$[\nabla_{W_1} D(V|\tilde{V})]^- = 2VV^TW_1,$$

and

$$[\nabla_{W_1} D(V|\tilde{V})]^+ = 2VH_2^TW_2^TW_1 + W_2H_2V^TW_1 + VV^TW_1W_1^TW_1 + W_1W_1^TVV^TW_1.$$

Similarly, the gradient wrt W_2 gives

$$[\nabla_{W_2} D(V|\tilde{V})]^- = VH_2^T$$

and

$$[\nabla_{W_2} D(V|\tilde{V})]^+ = 2W_1W_1^TVH_2^T + W_2H_2H_2^T.$$

Finally, the gradient wrt H_2 gives

$$[\nabla_{H_2} D(V|\tilde{V})]^- = W_2^TV$$

and

$$[\nabla_{H_2} D(V|\tilde{V})]^+ = 2W_2^TW_1W_1^TV + W_2^TW_2H_2.$$

B. Kullback Leiber divergence

The Kullback Leiber divergence gives us the problem,

$$\min_{W_1,W_2,H_2\geq 0} V(log(V) - log(\tilde{V})) + (V - \tilde{V}).$$

The gradient wrt W_1 gives

$$[\nabla_{W_1} D(V|\tilde{V})]^-_{i,j} = (ZV^T W_1)_{i,j} + (VZ^T W_1)_{i,j},$$

with $Z_{i,j} = (\frac{V}{W_1 W_1^T V + W_2 H_2})_{i,j}$. The positive part of the gradient is

$$[\nabla_{W_1} D(V|\tilde{V})]^+_{i,j} = \sum_k (W^T V)_{j,k} + (\sum_k V_{i,k})(\sum_a W_{a,j}).$$

Similarly, the gradient wrt W_2 gives

$$[\nabla_{W_2} D(V|\tilde{V})]^- = V H_2^T$$

and

$$[\nabla_{W_2} D(V|\tilde{V})]^+ = W_1 W_1^T V H_2^T + W_2 H_2 H_2^T.$$

Finally, the gradient wrt H_2 gives

$$[\nabla_{H_2} D(V|\tilde{V})]^- = W_2^T V$$

and

$$[\nabla_{H_2} D(V|\tilde{V})]^+ = 2W_2^T W_1 W_1^T V + W_2^T W_2 H_2.$$

C. Itakura Saito divergence

The Itakura Saito divergence gives us the problem,

$$\min_{W_1,W_2,H_2\geq 0} \frac{V}{\tilde{V}} - log(\frac{V}{\tilde{V}}) - 1.$$

The gradient wrt W_1 gives

$$[\nabla_{W_1} D(V|\tilde{V})]^-_{i,j} = (ZV^T W_1)_{i,j} + (VZ^T W_1)_{i,j},$$

with $Z_{i,j} = (\frac{V}{W_1 W_1^T V + W_2 H_2})_{i,j}$. The positive part of the gradient is

$$[\nabla_{W_1} D(V|\tilde{V})]^+_{i,j} = (\phi V^T W_1)_{i,j} + (V\phi^T W_1)_{i,j},$$

with

$$\phi_{i,j} = (\frac{I}{W_1 W_1^T V + W_2 H_2})_{i,j}.$$

and $I \in \mathbb{R}^{f\times t}; \forall i,j \quad I_{i,j} = 1$.

Similarly, the gradient wrt W_2 gives

$$[\nabla_{W_2} D(V|\tilde{V})]^- = V H_2^T$$

and

$$[\nabla_{W_2} D(V|\tilde{V})]^+ = W_1 W_1^T V H_2^T + W_2 H_2 H_2^T.$$

Finally, the gradient wrt H_2 gives

$$[\nabla_{H_2} D(V|\tilde{V})]^- = W_2^T V$$

and

$$[\nabla_{H_2} D(V|\tilde{V})]^+ = 2W_2^T W_1 W_1^T V + W_2^T W_2 H_2.$$

D. SAR and SIR results with the different dictionaries

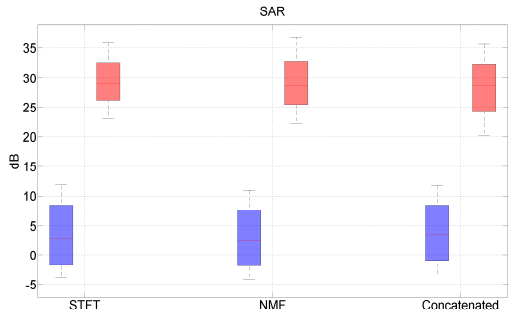

Fig. 11: SAR of harmonic (left bar)/percussive (right bar) estimated sources on the SiSec database with a STFT dictionary, a NMF dictionary and their concatenation.

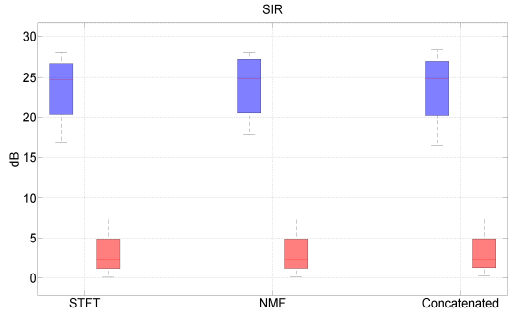

Fig. 12: SIR of harmonic (left bar)/percussive (right bar) estimated sources on the SiSec database with a STFT dictionary, a NMF dictionary and their concatenation.

E. Results on the sub database

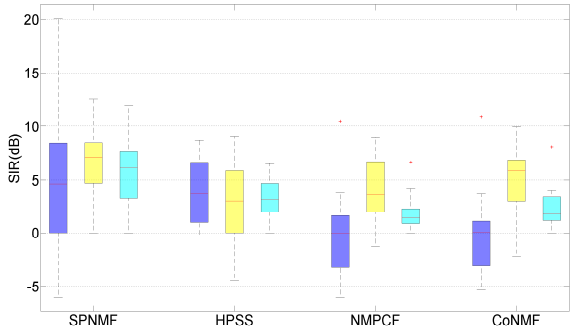

Fig. 13: SIR for percussive/left, harmonic/middle, mean/right separation results on the Electronic/Fusion songs for the four methods.

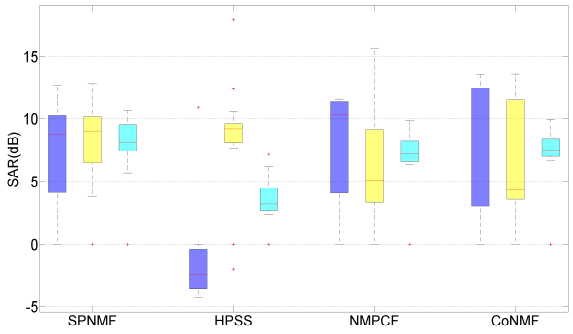

Fig. 14: SAR for percussive/left, harmonic/middle, mean/right separation results on the Electronic/Fusion songs for the four methods.

ACKNOWLEDGMENT

REFERENCES

[1] D. Lee and S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] S. Ewert and M. Müller, "Score-informed source separation for music signals," *Multimodal music processing*, vol. 3, pp. 73–94, 2012.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. of IEEE ICASSP*, vol. 1, 2007, pp. 65–68.

[4] R. Hennequin, R. Badeau, and B. David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *Proc. of DAFx*, 2010, pp. 246–253.

[5] J. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 564–575, 2010.

[6] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. of ISMIR*, 2007.

[7] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[8] X. Jaureguiberry, P. Leveau, S. Maller, and J. J. Burred, "Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation," in *Proc. of IEEE ICASSP*, 2011, pp. 5–8.

[9] C.-W. Wu and A. Lerch, "Drum transcription using partially fixed non-negative matrix factorization," in *Proc. of EUSIPCO*, 2008.

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

[11] F. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–17, 2014.

[12] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. of IEEE IJCNN*, 2008, pp. 1828–1832.

[13] Z. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," *IEEE Transactions on Neural Network.*, vol. 21, no. 5, pp. 734–749, 2010.

[14] E.Vincent, N.Bertin, and R.Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing.*, vol. 18, no. 3, pp. 528–537, 2010.

[15] C. Laroche, M. Kowalski, H. Papadopoulos, and G. Richard, "A structured nonnegative matrix factorization for source separation," in *Proc. of EUSIPCO*, 2015.

[16] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," in *Audio Engineering Society Convention 115*. Audio Engineering Society, 2003.

[17] L. Daudet, "A review on techniques for the extraction of transients in musical signals," in *Computer Music Modeling and Retrieval*. Springer, 2006, pp. 219–232.

[18] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Transactions on Audio, Speech, and Language Processing.*, vol. 16, no. 2, pp. 255–266, 2008.

[19] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *Proc. of IEEE ICASSP*, vol. 2, 2000, pp. II753–II756.

[20] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *Transactions on Audio, Speech, and Language Processing.*, vol. 20, no. 6, pp. 1759–1770, 2012.

[21] D. P. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[22] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proc. of EUSIPCO*, 2005, pp. 1–4.

[23] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of DAFx*, 2010.

[24] C.-L. Hsu, D. Wang, J.-S. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *Transactions on Audio, Speech, and Language Processing.*, vol. 20, no. 5, pp. 1482–1491, 2012.

[25] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. of EUSIPCO*, 2008.

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

[27] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," *Image Analysis*, pp. 333–342, 2005.

[28] D. Lee and S. Seung, "Algorithms for non-negative matrix factorization," *Proc. of NIPS*, pp. 556–562, 2001.

[29] O.Gillet and G.Richard, "Enst-drums: an extensive audio-visual database for drum signals processing." in *Proc. of ISMIR*, 2006, pp. 156–159.

[30] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. of IEEE ICASSP*, 2015.

[31] C. Févotte, N. Bertin, and JL.Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[32] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong, "The 2010 signal separation evaluation campaign : audio source separation," in *Proc. of LVA/ICA*, 2010, pp. 114–122.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, Language Process.*, vol. 14, pp. 1462–1469, 2006.

[34] R. M. Gray, A. Buzo, A. H. Gray Jr, and Y. Matsuyama, "Distortion measures for speech processing," *Transactions on Acoustics, Speech and Signal Processing.*, vol. 28, no. 4, pp. 367–376, 1980.

[35] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research," in *Proc. of ISMIR*, 2014.

[36] J. Driedger and M. Müller, "Tsm toolbox: Matlab implementations of time-scale modification algorithms," in *Proc. of DAFx*, 2014, pp. 249–256.