HARMONIC-PERCUSSIVE SOUND SEPARATION USING RHYTHMIC INFORMATION FROM NON-NEGATIVE MATRIX FACTORIZATION IN SINGLE-CHANNEL MUSIC RECORDINGS

F.J. Canadas-Quesada[1], *D. Fitzgerald*[2], *P. Vera-Candeas*[1], *N. Ruiz-Reyes*[1] *

[1] Telecommunication Engineering Department, Higher Polytechnic School of Linares, University of Jaen, Jaen, Spain
[2] Cork School of Music, Cork Institute of Technology, Cork, Ireland
fcanadas@ujaen.es, Derry.Fitzgerald@cit.ie, pvera@ujaen.es, nicolas@ujaen.es

ABSTRACT

This paper proposes a novel method for separating harmonic and percussive sounds in single-channel music recordings. Standard non-negative matrix factorization (NMF) is used to obtain the activations of the most representative patterns active in the mixture. The basic idea is to classify automatically those activations that exhibit rhythmic and non-rhythmic patterns. We assume that percussive sounds are modeled by those activations that exhibit a rhythmic pattern. However, harmonic and vocal sounds are modeled by those activations that exhibit a less rhythmic pattern. The classification of the harmonic or percussive NMF activations is performed using a recursive process based on successive correlations applied to the activations. Specifically, promising results are obtained when a sound is classified as percussive through the identification of a set of peaks in the output of the fourth correlation. The reason is because harmonic sounds tend to be represented by one valley in a half-cycle waveform at the output of the fourth correlation. Evaluation shows that the proposed method provides competitive results compared to other reference state-of-the-art methods. Some audio examples are available to illustrate the separation performance of the proposed method.

1. INTRODUCTION

Harmonic (pitched instruments) and percussive (drums) sound separation is still an unsolved problem in music signal processing and machine learning. It can be applied to Music Information Retrieval (MIR) in two ways. From a percussive point of view, it can enhance tasks such as, onset detection and tempo estimation. From a harmonic point of view, it can improve other tasks such as, score alignment, multi-pitch and melody estimation, chord detection or vocal extraction.

Rhythm can be considered of core importance in most music, and it is often provided by percussive sounds. Although there are some harmonic instruments (e.g., bass guitar) that show repetitive temporal behavior, in this paper, we assume that percussive sounds (repetitive) exhibit a more rhythmic pattern compared to harmonic sounds (non-repetitive). In this manner, rhythmic information could be useful to discriminate percussive and harmonic sounds in an acoustic mixture in the same manner that a non-trained listener can effortlessly discriminate between them.

In recent years, several approaches have been applied in the field of harmonic-percussive sound separation. Most of these approaches utilize the anisotropy of harmonic and percussive sounds, that is, percussive sounds have a structure that is vertically smooth in frequency, whereas harmonic sounds are temporally stable and have a structure that is horizontally smooth in time. Anisotropy is applied in a maximum a posteriori (MAP) framework in [1]. Furthermore, the concept of anisotropy is applied using median filtering assuming that harmonics and percussive onsets can respectively be considered outliers in a temporal or frequency slice of a spectrogram [2]. In [3], a non-negative matrix partial co-factorization is presented that forces some portion of bases to be associated with drums only. Kim et.al [4] develop an extension of non-negative matrix factorization (NMF) using temporal repeatability of the rhythmic sound sources among segments of a mixture. Canadas et.al [5] propose an unsupervised NMF integrating spectro-temporal features, such as anisotropic smoothness or time-frequency sparseness, into the factorization process. Kernel additive modeling (KAM) is used to separate sound sources assuming that individual time-frequency bins are close in value to other bins nearby in the spectrogram where nearness is defined through a source-specific proximity kernel [6]. Driedger et. al [7] enforce the components to be clearly harmonic or percussive by exploiting a third residual component that captures the sounds that lie in between the clearly harmonic and percussive sounds. Park and Lee [8] include sparsity and harmonicity constraints in a NMF approach which uses a generalized Dirichlet prior. In [9], the concept of percussive anisotropy is used assuming that the percussive chroma clearly shows an energy distribution which is approximately flat.

One of the main goals of this paper is to determine if only the use of rhythmic information can provide reliable information to discriminate between harmonic and percussive sources. In this paper, we propose a method to separate harmonic and percussive sounds only analyzing the temporal information contained in the activations obtained from non-negative matrix factorization. The basic idea is to classify automatically those activations that exhibit rhythmic (percussive) and non-rhythmic (harmonic and vocal) patterns, assuming that a rhythmic pattern models repetitive events typically shown by percussive sounds and a non-rhythmic pattern models non-repetitive events typically shown by harmonic or vocal sounds. Specifically, the proposed method uses a recursive process based on successive correlations applied to the NMF activations. As shown later, a percussive sound is characterized by a set of peaks at the output of the fourth correlation but a harmonic sound tends to be represented by one valley in a half-cycle waveform at the output of the fourth correlation. Some of the advantages of our proposal are (i) simplicity; (ii) no prior information about the spectral content of the musical instruments and (iii) no prior training.

The remainder of this paper is organized as follows. Section 2 introduces briefly the mathematical background related to the standard NMF. In Section 3, the proposed method is detailed. Sec-

* This work was supported by the Spanish Ministry of Economy and Competitiveness under Project TEC2015-67387-C4-2-R

tion 4 optimizes and evaluates the separation performance of the proposed method compared to reference state-of-the-art methods. Finally, conclusions and future work are presented in Section 5.

2. NON-NEGATIVE MATRIX FACTORIZATION

Standard (unconstrained) non-negative matrix factorization (NMF) [10] attempts to obtain a parts-based representation of the most representative objects in a matrix by imposing non-negative constraints. The basic concept of NMF can be expressed as $X_{F,T} \approx \hat{X}_{F,T} = W_{F,K} H_{K,T}$ where the mixture is modelled as the linear combination of K components. Specifically, $X_{F,T}$ represents the magnitude spectrogram of the mixture, where $f = 1, ..., F$ denotes the frequency bin and $t = 1, ..., T$ is the time frame, $\hat{X}_{F,T}$ is the estimated matrix, $W_{F,K}$ is the basis matrix whose columns are the basis functions (or spectral patterns) and $H_{K,T}$ is the activation matrix for the basis functions. The rank or number of components K is generally chosen such that $FK + KT \ll FT$ in order to reduce the dimensions of the data. The factorization is obtained by minimizing a cost function $D(X|\hat{X})$ defined as,

$$D(X|\hat{X}) = \sum_{f=1}^{F} \sum_{t=1}^{T} d(X_{f,t}|\hat{X}_{f,t}) \tag{1}$$

where $d(a|b)$ is a function of two scalar variables. In this work, we used the generalized Kullback-Leibler divergence $D(X|\hat{X}) = D_{\mathrm{KL}}(X|\hat{X})$ because it has been successfully applied in the field of sound source separation [11] [12] [13],

$$D_{\mathrm{KL}}(X|\hat{X}) = \left(X \odot \log\left(X \oslash \hat{X} \right) \right) - X + \hat{X} \tag{2}$$

where $\odot$ is the element-wise multiplication and $\oslash$ is the element-wise division.

The cost function $D_{\mathrm{KL}}(X|\hat{X})$ is minimized using an iterative algorithm based on multiplicative update rules and the non-negativity of the bases and the activations is ensured. In this manner, the multiplicative update rule for an arbitrary scalar parameter Z is computed as follows,

$$Z \leftarrow Z \odot \left(\left[\frac{\partial D_{\mathrm{KL}}(X|\hat{X})}{\partial Z} \right]^{-} \oslash \left[\frac{\partial D_{\mathrm{KL}}(X|\hat{X})}{\partial Z} \right]^{+} \right) \tag{3}$$

3. PROPOSED METHOD

The proposed method uses standard NMF to separate percussive sounds $x_p(t)$ from harmonic sounds $x_h(t)$ in single-channel music mixtures $x(t)$. The magnitude spectrogram X of a mixture $x(t)$, calculated from the magnitude of the short-time Fourier transform (STFT) using a N-sample hamming window $w(n)$ and a J-sample hop size, is composed of F frequency bins and T frames. We assume that percussive and harmonic sounds are mixed in an approximately linear manner, that is, $x(t) = x_p(t) + x_h(t)$ in the time domain or $X = X_p + X_h$ in the magnitude frequency domain. As a result, X can be factorized into two separated spectrograms, $\hat{X}_p$ (an estimated spectrogram only composed of percussive sounds) and $\hat{X}_h$ (an estimated spectrogram only composed of harmonic sounds),

$$X = X_p + X_h = \begin{bmatrix} W_p & W_h \end{bmatrix} \begin{bmatrix} H_p \\ H_h \end{bmatrix} \tag{4}$$

where W_p, H_p are the original percussive bases and activations; W_h, H_h are the original harmonic bases and activations. All the previous data are non-negative matrices. The flowchart of the proposed method is shown in Figure 1.

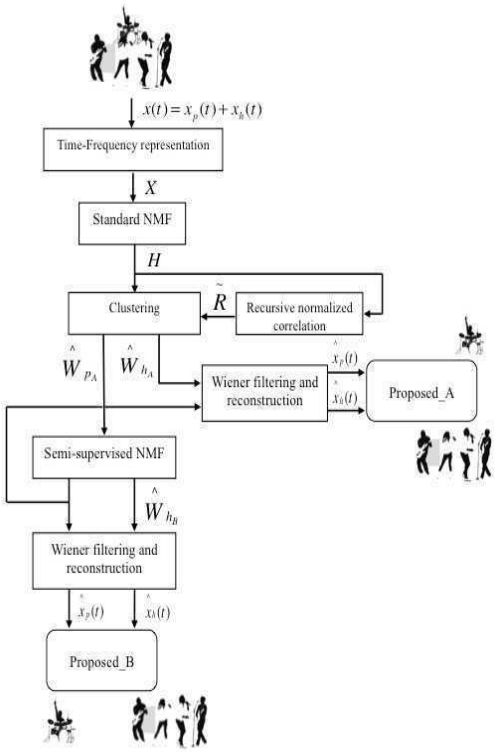


Figure 1: *Flowchart of the proposed method for the task of harmonic-percussive sound separation in single-channel music recordings.*

3.1. Obtaining activations

Standard NMF is applied to the magnitude spectrogram X using the cost function $D_{\mathrm{KL}}(X|\hat{X})$ previously mentioned in section 2. The update rules are defined as follows,

$$H \leftarrow H \odot \left(\left(W^T \left(X \oslash \hat{X} \right) \right) \oslash \left(W^T \mathbf{1}_{F,T} \right) \right) \tag{5}$$

$$W \leftarrow W \odot \left(\left(\left(X \oslash \hat{X} \right) H^T \right) \oslash \left(\mathbf{1}_{F,T} H^T \right) \right) \tag{6}$$

where W and H are initialized as random positive matrices, $\mathbf{1}_{F,T}$ represents a matrix of all-ones composed of F rows and T columns and ^{T} is the transpose operator. Note that in this paper we normalise each i^{th} basis function using the L^2-norm, that is, $\tilde{W}_i$=$\frac{W_i}{\|W_i\|_2}$, being $\left\|\tilde{W}_i\right\|_2$=1.0.

Standard NMF can only ensure convergence to local minima, which enables the reconstruction of the mixture but cannot distinguish by itself if the i^{th} component represents a percussive or harmonic sound.

3.2. Recursive normalized correlation and clustering

Our main contribution attempts to discriminate harmonic and percussive sounds only analyzing the activations H. The basic idea is to classify automatically those activations that exhibit rhythmic and non-rhythmic patterns. We assume that a rhythmic pattern models repetitive events typically associated with percussive sounds. A non-rhythmic pattern is assumed as a non-repetitive event typically shown by harmonic or vocal sounds.

We develop a recursive process, based on the normalized unbiased correlation $\tilde{R}^L(\tau)$ with order L, to identify rhythmic and non-rhythmic patterns. The normalized unbiased correlation $\tilde{R}^L(\tau)$ is computed using as input the signal $I(t)$ as shown in eq (7). We define the order L as the number of times that the normalized unbiased correlation is computed using the recursive process. In this manner, $I(t)$=$H(t)$ for L=0, $I(t)$=$\tilde{R}^0(\tau)$ for L=1, $I(t)$=$\tilde{R}^1(\tau)$ for L=2, etc. In a recursive way, the output of the current order L will be the input of the next order $L+1$. Specifically, the normalized unbiased correlation $\tilde{R}^L(\tau)$ is computed in eq. (8),

$$R^L(\tau) = \frac{1}{T-\tau} \sum_{t=0}^{T-1-\tau} I(t)I(t+\tau), \tau = 0, 1, 2, ..., T-1 \quad (7)$$

$$\tilde{R}^L(\tau) = \frac{R^L(\tau)}{\left\|R^L(\tau)\right\|_2} \quad (8)$$

The analysis of $\tilde{R}^L(\tau)$ indicates that $\tilde{R}^4(\tau)$ provides reliable information to discriminate harmonic and percussive sounds as can be observed in Figure 2 and Figure 4. Figure 2 and Figure 4 show the matrix H of activations of a music excerpt composed of harmonic and percussive sounds. It can be observed that the components 4, 5, 9, 14 and 17 of Figure 2 and the components 14, 15, 17 and 18 of Figure 4 represents predominant percussive sounds modeled by rhythmic patterns. Both Figure 3 and Figure 5 show that $\tilde{R}^L(\tau)$ is still showing a set of peaks even as the order L increases when a percussive sound is analyzed. However, this does not occur when analyzing harmonic sounds since these tend to be represented using only one valley in a half-cycle waveform when the order L is increased. It can be observed that $\tilde{R}^4(\tau)$ is optimal (see section 4.4) to discriminate between percussive and harmonic sounds because $\tilde{R}^4(\tau)$ clearly shows a set of peaks considering percussive sounds and only one valley in a half-cycle waveform considering harmonic sounds. As a result, $\tilde{R}^{L<4}(\tau)$ could model harmonic sounds as percussive sounds because $\tilde{R}^{L<4}(\tau)$ can represent harmonic sounds with more than one peak as shown in Figures 3(b), 3(d), 3(f) and 3(h) and Figures 5(b), 5(d), 5(f) and 5(h). However, $\tilde{R}^{L>4}(\tau)$ could model percussive sounds as harmonic sounds because $\tilde{R}^{L>4}(\tau)$ tends to remove most of the peaks as shown in Figure 3(k) and Figure 3(m) and Figure 5(k) and Figure 5(m).

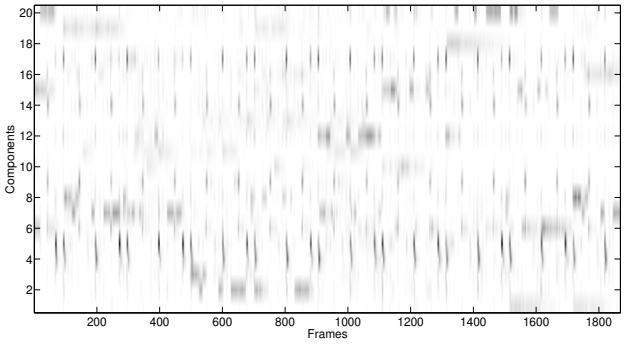


Figure 2: *Activations H from the excerpt 'Hotel California' (Table 2), using K=20 components.*

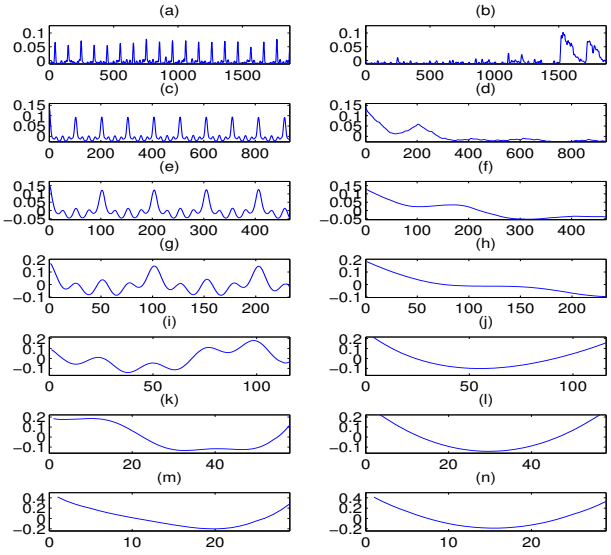


Figure 3: *The column of the left represents the percussive component 14 of Figure 2: (a) $\tilde{R}^0(\tau)$, (c) $\tilde{R}^1(\tau)$, (e) $\tilde{R}^2(\tau)$, (g) $\tilde{R}^3(\tau)$, (i) $\tilde{R}^4(\tau)$, (k) $\tilde{R}^5(\tau)$, (m) $\tilde{R}^6(\tau)$. The column of the right represents the harmonic component 1 of Figure 2: (b) $\tilde{R}^0(\tau)$, (d) $\tilde{R}^1(\tau)$, (f) $\tilde{R}^2(\tau)$, (h) $\tilde{R}^3(\tau)$, (j) $\tilde{R}^4(\tau)$, (l) $\tilde{R}^5(\tau)$, (n) $\tilde{R}^6(\tau)$.*

Based on the above observation, we use $\tilde{R}^4(\tau)$ as the basis for automatically discriminating between percussive and harmonic sounds. A component, obtained from NMF decomposition, is classified as percussive if a set of N_p or more peaks are found at the output of the $\tilde{R}^4(\tau)$. Preliminary results indicated that the best separation performance was obtained when $N_p \geq 2$. In any other case, a component is classified as harmonic.

We have developed two approaches based on the classification of the rhythmic activations as shown in Figure 1. In the first approach called Proposed_A, $\hat{W}_{p_A}$ and $\hat{H}_{p_A}$ represent the estimated bases and activations classified as percussive. However, $\hat{W}_{h_A}$ and $\hat{H}_{h_A}$ represent the estimated bases and activations clas-

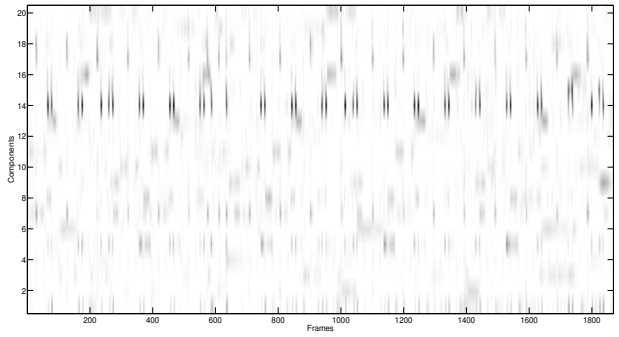


Figure 4: *Activations H from the excerpt 'So lonely' (Table 1), using K=20 components.*

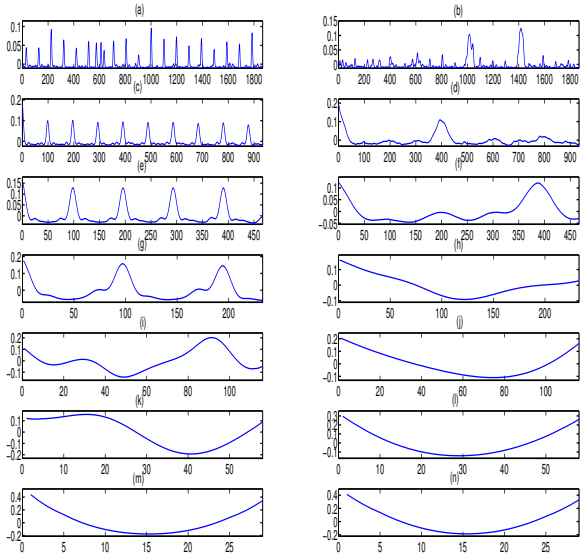


Figure 5: *The column of the left represents the percussive component 17 of Figure 4: (a) $\tilde{R}^0(\tau)$, (c) $\tilde{R}^1(\tau)$, (e) $\tilde{R}^2(\tau)$, (g) $\tilde{R}^3(\tau)$, (i) $\tilde{R}^4(\tau)$, (k) $\tilde{R}^5(\tau)$, (m) $\tilde{R}^6(\tau)$. The column of the right represents the harmonic component 2 of Figure 4: (b) $\tilde{R}^0(\tau)$, (d) $\tilde{R}^1(\tau)$, (f) $\tilde{R}^2(\tau)$, (h) $\tilde{R}^3(\tau)$, (j) $\tilde{R}^4(\tau)$, (l) $\tilde{R}^5(\tau)$, (n) $\tilde{R}^6(\tau)$.*

sified as harmonic. Specifically, $\hat{W}_p = \hat{W}_{p_A}$ and $\hat{H}_p = \hat{H}_{p_A}$, $\hat{W}_h = \hat{W}_{h_A}$, $\hat{H}_h = \hat{H}_{h_A}$

In the second approach called Proposed_B, a semi-supervised NMF, based on the Kullback-Liebler divergence, is used. The estimated percussive bases $\hat{W}_{p_B}$ are fixed to $\hat{W}_p$=$\hat{W}_{p_B}$=$\hat{W}_{p_A}$ and not updated. However, the estimated harmonic bases and the estimated percussive and harmonic activations are initialized $\hat{W}_h$=$\hat{W}_{h_B}$=$\hat{W}_{h_A}$, $\hat{H}_p$=$\hat{H}_{p_B}$=$\hat{H}_{p_A}$ and $\hat{H}_h$=$\hat{H}_{h_B}$=$\hat{H}_{h_A}$ and updated in the factorization process.

3.3. Reconstruction and Wiener filtering

Performing each approach, the separated percussive and harmonic signals $\hat{x}_p(t)$, $\hat{x}_h(t)$ are synthesized using the magnitude spectrogram, that is, $\hat{X}_p = \hat{W}_p\hat{H}_p$ and $\hat{X}_h = \hat{W}_h\hat{H}_h$.

If the power spectral density (PSD) of the estimated signals are denoted as $|\hat{X}_p|^2$ and $|\hat{X}_h|^2$, respectively, then each ideally estimated source $\hat{x}_p(t)$ or $\hat{x}_h(t)$ can be estimated from the mixture $x(t)$ using a generalized time-frequency mask over the STFT domain. To ensure that the reconstruction process is conservative, a Wiener filtering has been used as in [5]. In this manner, a percussive or harmonic Wiener mask represents the relative percussive or harmonic energy contribution of each type of sound with respect to the energy of the mixture defined as follows,

$$\hat{X}_{\mathrm{p}} = \left(|\hat{X}_{\mathrm{p}}|^2 \oslash \left(|\hat{X}_{\mathrm{p}}|^2 + |\hat{X}_{\mathrm{h}}|^2 \right) \right) \odot X \tag{9}$$

$$\hat{X}_{\mathrm{h}} = \left(|\hat{X}_{\mathrm{h}}|^2 \oslash \left(|\hat{X}_{\mathrm{p}}|^2 + |\hat{X}_{\mathrm{h}}|^2 \right) \right) \odot X \tag{10}$$

The estimated percussive and harmonic signals $\hat{x}_p(t)$, $\hat{x}_h(t)$ are obtained computing the inverse overlap-add STFT from the final estimated percussive and harmonic magnitude spectrograms $\hat{X}_p$, $\hat{X}_h$ using the phase spectrogram of the mixture.

4. EXPERIMENTAL RESULTS

4.1. Data and metrics

Evaluation has been performed using two databases DBO and DBT. Both databases are composed of single-channel real-world music excerpts taken from the Guitar Hero game [14] [15] as can be seen in Table 1 and Table 2. Each excerpt has a duration about 30 seconds and it was converted from stereo to mono and sampled at f_s=16 kHz.

The database DBO has been used in the optimization and the database DBT has been used in testing. Note that the database used in the optimization is not the same as that used in the testing to validate the results. The subscript $_{ph}$ is related to percussive and harmonic instrumental sounds without adding the original vocal sounds. In this case, each percussive signal $x_p(t)$ is composed of percussive sounds (drums) and each harmonic signal $x_h(t)$ is composed of harmonic instrumental sounds. However, the subscript $_{phv}$ is related to percussive, harmonic and vocal sounds. As a result, each percussive signal $x_p(t)$ is composed of percussive sounds (drums) and each harmonic signal $x_h(t)$ is composed of harmonic and vocal sounds.

Table 1: Title and artist of the excerpts of the databases DBO_{ph} and DBO_{phv}

TITLE	ARTIST
Are you gonna go my way	Lenny Kravitz
Feel the pain	Dinosaur Jr
Kick out the James	MC5's Wayne Krame
One way or another	Blondie
In my place	Coldplay
Livin' on a prayer	Bon Jovi
No one to depend on	Santana
So lonely	The police
Song 2	Blur

Table 2: Title and artist of the excerpts of the databases DBT_{ph} and DBT_{phv}

TITLE	ARTIST
Hollywood Nights	Bob Seger & The Silver Bullet Band
Hotel California	Eagles
Hurts So Good	John Mellencamp
La Bamba	Los Lobos
Make It Wit Chu	Queens Of The Stone Age
Ring of Fire	Johnny Cash
Rooftops	Lost prophets
Sultans of Swing	Dire Straits
Under Pressure	Queen

The assessment of the performance of the proposed method has been performed using the metrics Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR) and Source to Artifacts Ratio (SAR) [16] [17] which are widely used in the field of sound source separation. Specifically, SDR provides information on the overall quality of the separation process. SIR is a measure of the presence of percussive sounds in the harmonic signal and vice versa. SAR provides information on the artifacts in the separated signal from separation and/or resynthesis. Higher values of these ratios indicate better separation quality. More details can be found in [16].

4.2. Setup

An initial evaluation has been performed taking into account the computation of the STFT in order to optimize the frame size N = (1024, 2048 and 4096 samples) using the sampling rate f_s previously mentioned. Preliminary results indicated that the best separation performance was achieved using (N,J)=(2048,256) samples.

A random initialization of the matrices W and H was used and the convergence of the NMF decomposition was evaluated using $Niter$ iterations. Due to the fact that standard NMF is not guaranteed to find a global minimum, the performance of the proposed method depends on the initial values W and H obtaining different results. For this reason, we have repeated three times for each excerpt and the results in the paper are averaged values.

4.3. State-of-the-art methods

Two reference state-of-the-art percussive and harmonic separation methods have been used to evaluate the proposed method: HPSS [1] and MFS [2]. These were both implemented for the evaluation of this paper. The ideal separation, called Oracle, is provided to better compare the quality of the proposed methods. The Oracle separation has been computed using the ideal soft masks extracted from the original percussive and harmonic signals applied to the input mixture.

4.4. Optimization

An optimization of the parameters K, $Niter$ and L is performed in the databases DBO_{ph} and DBO_{phv} as shown in Figure 6 and Figure 7. For this purpose, a hyperparametric analysis is applied to each parameter of the proposed method as occurs in [5]. In this work, K= (5, 10, 20, 30, 50, 100, 150, 200), $Niter$=(10, 20, 30, 50, 100, 150) and L=(0, 1, 2, 3, 4, 5, 6).

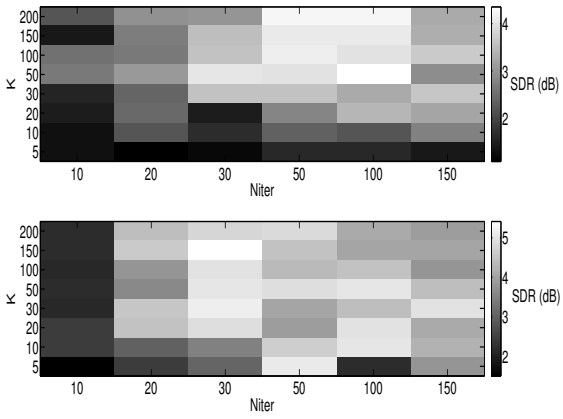


Figure 6: *Optimization of the parameters K and $Niter$ (using L=4) jointly averaging percussive and harmonic SDR. (top) evaluating the database DBO_{ph}, (bottom) evaluating the database DBO_{phv}*

Figure 6(top) shows that the parameters K=50 and $Niter$=100 maximizes the average percussive and harmonic SDR only considering mixtures composed of percussive and harmonic sounds without vocal sounds. It can be observed that using K<20 and $Niter$<30 provides the worst separation performance of the method. Results suggest that using a small number of components does not allow sufficient separation of repetitive and non-repetitive elements, thereby causing poor separation quality. Further, a small number of iterations does not allow to converge the NMF decomposition.

Figure 6(bottom) shows that K=150 and $Niter$=30 maximize the average percussive and harmonic SDR only considering mixtures composed of percussive, harmonic and vocal sounds. Figure 6(bottom) shows that a higher number of components is necessary to obtain the highest SDR. The effect of adding vocal sounds indicates the presence of a higher variety of spectral patterns so the proposed method needs a higher number of components to represent the mixture adequately..

Using the optimal parameters in the databases DBO_{ph} and DBO_{phv}, the optimization of the parameter L is shown in Figure 7. Comparing the separation performance of the parameter L, L=4 provides the highest robustness to discriminate harmonic and percussive sounds evaluating audio mixtures with or without vocal sounds. The principal reason for this is that $\tilde{R}^4(\tau)$ retains sufficient peaks in the repetitive basis functions to allow discrimination from the non-repetitive basis functions which tend to have a single valley at L=4.

4.5. Results

Figure 8(a) and Figure 8(b) show SDR, SIR and SAR results evaluating the databases DBT_{ph} and DBT_{phv} for the proposed method and the reference state-of-the-art methods. Each box represents nine data points, one for each excerpt of the test database. The lower and upper lines of each box show the 25^{th} and 75^{th} percentiles for the database. The line in the middle of each box represents the median value of the dataset. The left (blue), center (red) and right (black) boxes are related to the estimated percussive, har-

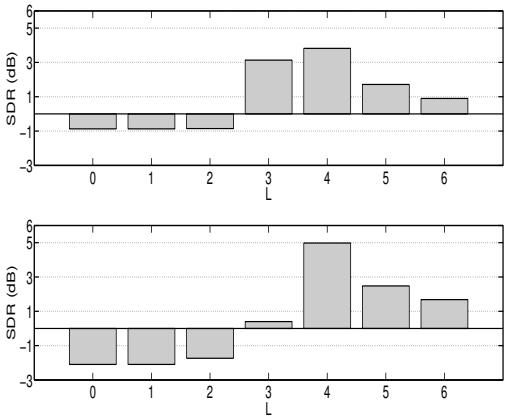


Figure 7: *Optimization of the parameters L jointly averaging percussive and harmonic SDR. (top) evaluating the database DBO_{ph} using the optimal parameters K=50 and $Niter$=100; (bottom) evaluating the database DBO_{phv} using the optimal parameters K=150 and $Niter$=30*

monic and average between percussive and harmonic signals.

Figure 8(a) indicates that MFS and the proposed method outperform the separation performance of HPSS in percussive and harmonic SDR but HPSS can be considered as competitive method. Although MFS and Proposed_B exhibit a similar behavior in percussive SDR, MFS (SDR = 5.9dB) is slightly better than Proposed_B (SDR = 5.0dB) in harmonic SDR. However, Proposed_B (SIR = 9.8dB) is slightly better than MFS (SIR = 9.1dB) considering the average between percussive and harmonic SIR. HPSS obtains the highest percussive SIR at the expense of introducing a high amount of artifacts, as shown by having the worst percussive and harmonic SAR. This does not occur with either MFS or the proposed methods. Further, HPSS loses most of the transients associated with the beginning of the harmonic sounds, thereby obtaining low harmonic SDR compared to the other methods. Although Proposed_A and Proposed_B show similar separation performance, it seems that Proposed_B is slightly better than Proposed_A. This can be explained because the semi-supervised NMF, initialized with the harmonic bases from Proposed_A, tends to converge to a better solution, obtaining a higher quality reconstruction of the estimated harmonic signals. Informal listening tests suggest that the proposed methods achieves higher polyphonic richness of the harmonic sounds compared to HPSS and MFS because it is able to capture most of the onsets of the harmonic sounds, such as onsets played by bass guitar or lead guitar. Nevertheless, a weakness of the proposed method is that it classifies as a percussive sound those harmonic sounds, e.g., bass guitar, that exhibit a very strong rhythmic pattern, especially if the bass guitar is playing a repeated note which has been factorized together with a part of a percussive sound in the same NMF component.

Figure 8(b) shows the separation performance for the database DBT_{phv}. Comparing with Figure 8(a), it can be observed that the addition of vocal sounds significantly worsens the SDR, SIR and SAR obtained by HPSS and MFS. However, the proposed methods still perform strongly, even with the addition of vocals. Results indicate that Proposed_A and Proposed_B show similar or even better SDR, SIR and SAR results compared to their separation performance in Figure 8(a). The evaluation metrics indicate that the proposed methods offer a more robust performance, especially when comparing their percussive and harmonic SDR and SIR results. Specifically, the proposed methods remove most of the vocal sounds from the estimated percussive signal, while HPSS and MFS do not. In some excerpts, this fact implies that the vocal sounds are unintelligible in the estimated harmonic signal. The promising behavior of the proposed methods with respect to vocal sounds can be explained by the fact that vocal sounds tend to exhibit less repetition, both in terms of melody and in terms of modulations than other sound sources.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a novel approach for separating harmonic and percussive sounds using only temporal information extracted from activations by means of non-negative matrix factorization. The basic idea is to classify automatically those activations that exhibit rhythmic and non-rhythmic patterns. We assume that a rhythmic pattern models repetitive events which are typically associated with percussive sounds. However, a non-rhythmic pattern models non-repetitive events typically associated with harmonic or vocal sounds. Some of the advantages of this approach are (i) simplicity; (ii) no prior information about the spectral content of the musical instruments and (iii) no prior training.

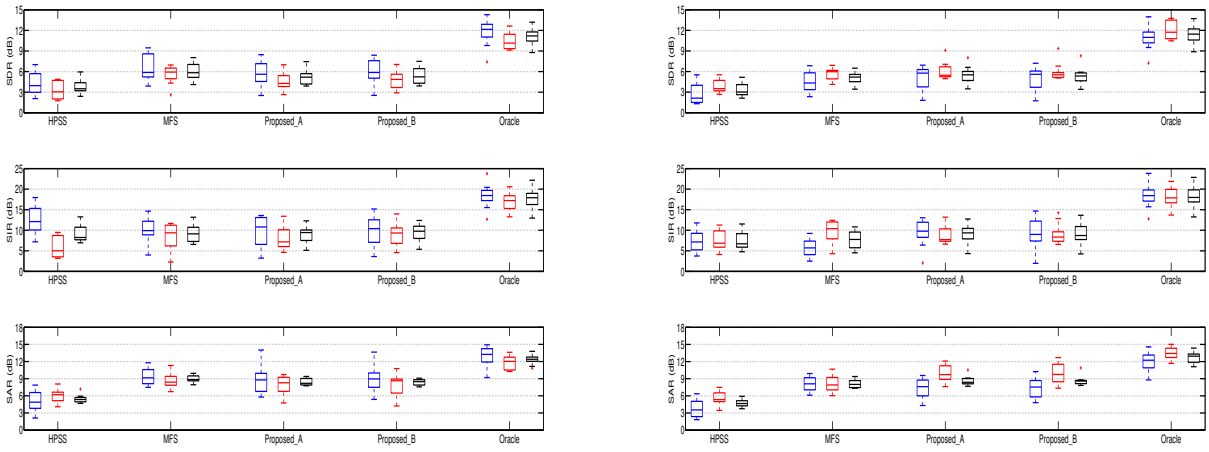
Evaluating instrumental mixtures without vocals, results indicate that the proposed methods obtain promising audio separation. The performance of Proposed_B is slightly better than Proposed_A because the update of the semi-supervised NMF allows convergence to a better solution which provides higher quality reconstruction of the estimated harmonic signals. Moreover, our approach obtains higher polyphonic richness of the harmonic sounds because it captures most of the onsets of the harmonic sounds. However, a weakness of the proposed method is that highly repetitive harmonic instruments can occasionally be classified as percussive.

Evaluating instrumental mixtures containing vocals, results show that the proposed method gives a more robust performance both in percussive and harmonic SDR and SIR compared to the reference state-of-the-art methods.The proposed method extracts most of the vocal sounds from the estimated percussive signal unlike the other reference state-of-the-art methods evaluated.

Future work will be focused on three directions. Firstly, we will try to improve the audio quality of the estimated sources using another time-frequency representation that provides a better resolution in the low frequency bands. Secondly, we will attempt to remove residual harmonic sounds that have been factorized in percussive NMF components. Thirdly, we will address how to separate harmonic sounds, e.g. bass guitar, that show temporally repetitive characteristics from percussive sounds which have been factorized in the same NMF component, by looking into other ways of extracting the periodicity of the activations.

6. REFERENCES

[1] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2008, pp. 25–29.



(a) Database DBT_{ph}.

(b) Database DBT_{phv}.

Figure 8: SDR, SIR and SAR results of the proposed methods and the state-of-the-art methods. The left (blue), center (red) and right (black) boxes are related to the estimated percussive, harmonic and average between percussive and harmonic signals.

[2] D. Fitzgerald, "Harmonic-percussive separation using median filtering," in *Proceedings of Digital Audio Effects (DAFX)*, 2010.

[3] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 1942–1945.

[4] M. Kim, J. Yoo, K. Kang, and S. Choi, "Blind rhythmic source separation: Nonnegativity and repeatability," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[5] F. Canadas, P. Vera, N. Ruiz, J. Carabias, and P. Cabanas, "Percussive-harmonic sound separation by non-negative matrix factorization with smoothness-sparseness constraints," *Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 26, pp. 1–17, 2014.

[6] D. Fitzgerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic-percussive separation using kernel additive modelling," in *25th IET Irish Signals and Systems Conference*, 2014.

[7] J. Driedger, M. Muller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[8] J. Park and K. Lee, "Harmonic-percussive source separation using harmonicity and sparsity constraints," in *16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[9] F. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, A. Munoz-Montoro, and F. Bris-Penalver, "A method to separate musical percussive sounds using chroma spectral flatness," in *The First International Conference on Advances in Signal, Image and Video Processing (SIGNAL)*, 2016.

[10] D. Lee and S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of Advances in Neural Inf. Process. System*, 2000, pp. 556–562.

[11] S. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

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

[13] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2096–2107, 2013.

[14] Activision Neversoft, "Guitar hero 5," http://gh5.guitarhero.com, September 2009.

[15] Activision Neversoft, "Guitar hero world tour," http://worldtour.guitarhero.com/us/, November 2008.

[16] C. Fevotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide - revision 2.0," in *Technical report 1706, IRISA*, 2005.

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