Canadas-Quesada *et al. EURASIP Journal on Audio, Speech, and Music Processing* 2014, **2014**:26
http://asmp.eurasipjournals.com/content/2014/1/26

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints

Francisco Jesus Canadas-Quesada[1*], Pedro Vera-Candeas[1], Nicolas Ruiz-Reyes[1], Julio Carabias-Orti[2] and Pablo Cabanas-Molero[1]

Abstract

In this paper, unsupervised learning is used to separate percussive and harmonic sounds from monaural non-vocal polyphonic signals. Our algorithm is based on a modified non-negative matrix factorization (NMF) procedure that no labeled data is required to distinguish between percussive and harmonic bases because information from percussive and harmonic sounds is integrated into the decomposition process. NMF is performed in this process by assuming that harmonic sounds exhibit spectral sparseness (narrowband sounds) and temporal smoothness (steady sounds), whereas percussive sounds exhibit spectral smoothness (broadband sounds) and temporal sparseness (transient sounds). The evaluation is performed using several real-world excerpts from different musical genres. Comparing the developed approach to three current state-of-the art separation systems produces promising results.

1 Introduction

The separation of percussive and harmonic sounds remains a challenging problem in music research. Percussive sound pertains to drum instruments, whereas the harmonic sound pertains to pitched instruments. We develop a method to separate monaural music signals (for which spatial information is unavailable), motivated by the large number of one-channel music recordings such as live performances or old recordings (from before the 1960s). In a musical context, a listener can effortlessly distinguish between percussive and harmonic sounds; therefore, these two types of sounds must have significantly different characteristics. Ono et al. [1,2] used Harmonic/Percussive Sound Separation (HPSS) to separate harmonic and percussive sounds by exploiting the anisotropy of harmonic and percussive sounds in a maximum a posteriori (MAP) framework. The authors considered a spectrogram, assuming anisotropic smoothness [1], i.e., 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.

Over the last decade, several approaches have been developed to separate percussive/harmonic sounds from monaural polyphonic music [3-6]. The method developed here offers the following advantages over the method outlined in [3]: (i) The method is robust for various sources and not only for flat spectrum sources, (ii) threshold choices are not required, (iii) hand-tuning is not necessary, (iv) the method is quite fast (e.g., the developed method can factorize an input signal lasting 30 s in approximately 18 s). Unlike the methods given in [4,5], no labelled data is required by the proposed method because the percussive/harmonic information is obtained from the spectro-temporal features used in the factorization stage. Other recently published state-of-the-art techniques are presented in [7,8]. In [7], anisotropy [1] is used in Median Filtering-based Separation (MFS) under two assumptions: the harmonics are considered to be outliers in a temporal slice that contains a mixture of percussive and pitched instruments, and the percussive onsets are considered to be outliers in a frequency slice. A median operator is used to remove these outliers because median filtering has been used extensively in image processing for removing salt and pepper noise from images [9]. In this manner, the

*Correspondence: fcanadas@ujaen.es
[1]Telecommunication Engineering Department, University of Jaen, Linares, Jaen, Spain
Full list of author information is available at the end of the article

extraction of percussive sounds can be seen as the removal of outliers (overtones from harmonic sounds) in a time frame of a spectrogram while the extraction of harmonic sounds can be seen as the removal of outliers (onsets from percussive sounds) in a frequency bin of a spectrogram. In [8], drum source separation is performed using non-negative matrix partial co-factorization (NMPCF). In NMPCF, the input spectrogram and a drum-only matrix (which is picked up from *a priori* drum recordings) are simultaneously decomposed. The shared basis vectors in this co-factorization are associated with the drum characteristics, which are used to extract drum-related components from the musical signals.

In this paper, we develop a percussive/harmonic sound separation approach, which we apply to monaural instrumental music (the singing voice is not considered). We do not consider vocals for the following reasons: (i) Voiced vocals can show harmonic features modelled by the developed method (e.g., smoothness in time for sustained sounds and sparseness in frequency for harmonic sounds) but can also exhibit harmonic features that are not modelled by the developed method (e.g., non-smoothness in time, as in the vibrato effect); (ii) vocals may be voiced (harmonic sounds) and unvoiced (percussive sounds), and the developed method has not been designed to distinguish between percussive music instruments and unvoiced vocals. The novelty of this work lies in modeling a percussive/harmonic mixture signal using a modified non-negative matrix factorization (NMF) approximation that can automatically distinguish between percussive and harmonic bases. That is, we decompose the mixture signal using an objective function to integrate spectro-temporal features, anisotropic smoothness and sparseness into the decomposition process. Anisotropic smoothness is related to the difference in the directions of continuity between the spectrograms of harmonic and percussive sounds. The spectrograms of harmonic sounds are quasi-stationary and are therefore typically smooth in time, whereas the spectrograms of percussive sounds are impulsive and are typically smooth in frequency [1,10]. However, anisotropic sparseness is also related to the difference in the directions of sparseness between the spectrograms of harmonic and percussive sounds. The spectrograms of harmonic sounds are typically sparse in frequency, as in narrowband sounds, whereas the spectrograms of percussive sounds are impulsive and are therefore typically sparse in time. These features enable us to model harmonic sounds using the sparseness in frequency (for spectral peaks) and smoothness in time (for amplitudes that vary slowly in time), whereas percussive sounds can be modelled using the smoothness in frequency (i.e., the energy slowly decreases in frequency) and the sparseness in time (i.e., most of the signal energy is concentrated over short time intervals), as seen in Figure 1. Therefore, the signal spectrogram can be reconstructed as the sum of two different spectrograms that are characterized by specific percussive/harmonic bases and gains. Our formulation does not require information about the number of active sound sources neither prior knowledge about the instruments nor supervised training to classify the bases.

The developed approach is practically useful in the field of audio engineering applications for music information retrieval, where the percussive/harmonic separation task can be used as a preprocessing tool. The extraction of a harmonic sound source can be used to enhance music transcription [11] and chord detection [12]. Extracting a percussive sound source can also enhance onset detection [2]. Extracting both harmonic and percussive sound sources is useful for remixing and for audio to score alignment [13].

We implement an unsupervised approach in which imposed smoothness and sparseness constraints are used to automatically discriminate between percussive and harmonic signals in a NMF framework. Our specific contribution is the inclusion of sparseness criteria in a NMF framework for percussive/harmonic separation. Compared to methods that require some training (semi-supervised or supervised), our approach provides a more robust source-to-distortion ratio (as we will show in Section 3) because the separation process does not depend on a supervised training. In Section 3, we show that the developed method produces promising results in comparison with two unsupervised (i.e., untrained) approaches (HPSS and MFS) and a supervised (i.e., trained) approach (NMPCF) dedicated to percussive/harmonic separation.

The remainder of this paper is organized as follows: In Section 2, we describe our novel method. In Section 3, the results are evaluated and compared. Finally, conclusions and future work are presented in Section 4.

2 Developed method

Non-negative matrix factorization (NMF) has been widely used in the field of digital image processing [14-17] in recent years and has also been successfully applied to music analysis [18,19]. Following [5,6,8], we apply NMF to percussive/harmonic separation, motivated in part by the aforementioned promising results.

Lee and Seung [20] developed standard NMF, a technique for multivariate data analysis in which an input magnitude spectrogram, represented by a matrix X, is decomposed into the product of two non-negative matrices W and H:

$$X \approx WH \tag{1}$$

where the i-th column of matrix W is a frequency basis that represents the spectral pattern of a component that is active in the spectrogram. Additionally, the i-th row

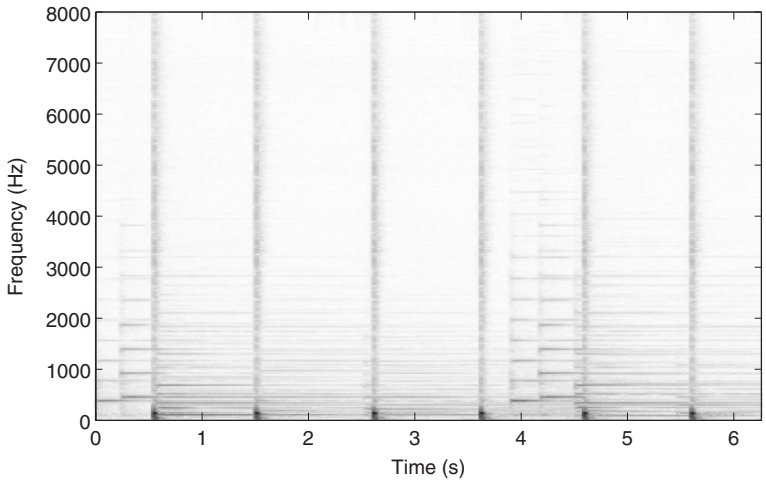

Figure 1 Magnitude spectrogram of an approximately 6-s-long excerpt of monaural mixture signal composed of percussive and harmonic sounds. The percussive sounds form vertical lines because of their smoothness in frequency and sparseness in time; the harmonic sounds form horizontal lines because of their sparseness in frequency and smoothness in time; the grey level represents the energy of each frequency.

of matrix H is a temporal gain or activation and represents the time interval over which the i-th frequency basis is active. Standard NMF cannot be used to distinguish between a percussive or harmonic frequency basis. Standard NMF can only ensure convergence to local minima, which enables the reconstruction of the signal spectrogram but cannot discriminate between percussive and harmonic frequency bases.

For clarity, the term *source* refers to a musical instrument, and the term *component* is used to select a specific frequency basis.

2.1 Signal representation

The magnitude spectrogram X of a music signal $x(t)$ is composed of T frames, F frequency bins and a set of time-frequency units $X_{f,t}$. Each $X_{f,t}$ is defined by the f-th frequency bin at the t-th frame and is calculated from the magnitude of the short-time Fourier transform (STFT) using a N-sample Hamming window $w(n)$ and a time shift J. A normalization process is necessary to adequately perform percussive/harmonic separation, for which the algorithm is independent of the norm of the input signal. Thus, the normalized magnitude spectrogram X_{n_β} is computed taking into account its dependence on the number of frames, the number of frequency bins and the constant β used in the β-divergence cost (see subsection 2.2.1).

$$X_{n_\beta} = \frac{X}{\left(\frac{\sum_{f=1}^{F}\sum_{t=1}^{T} X_{f,t}^{\beta}}{FT}\right)^{\frac{1}{\beta}}} \tag{2}$$

2.2 A modified non-negative matrix factorization for percussive/harmonic separation

Our formulation attempts to overcome the primary problem of the standard NMF approach by distinguishing between percussive and harmonic bases in the factorization process. For this purpose, an objective function is defined to decompose a mixture spectrogram X_{n_β} into two separate spectrograms, X_P (a percussive spectrogram) and X_H (a harmonic spectrogram) [5,6,8]. Each separated spectrogram exhibits specific spectro-temporal features for percussive or harmonic sounds. The factorization model is given in Equation 3 below:

$$X_{n_\beta} \approx X_P + X_H = W_{P_{F,R_p}} H_{P_{R_p,T}} + W_{H_{F,R_h}} H_{H_{R_h,T}} \tag{3}$$

where X_P, X_H, W_P, H_P, W_H and H_H are non-negative matrices. The parameter R_p denotes the number of percussive components, and the parameter R_h denotes the number of harmonic components used in the factorization process.

Next, we detail the decomposition process. This process adapts the concept of anisotropy which was initially appropriated by [1,2] to a NMF framework. Anisotropy is used to estimate W_P, H_P, W_H and H_H by minimizing a global objective function that depends on a β-divergence cost, a percussive cost and a harmonic cost.

2.2.1 β-divergence cost

Two different spectrograms X_P and X_H are constructed as to minimize the β-divergence cost $d_\beta(x|y)$ [21] compared to the input (known and fixed) normalized spectrogram X_{n_β}. The Euclidean distance ($\beta = 2$), the Kullback-Leibler (KL) divergence ($\beta = 1$) and the Itakura-Saito (IS) divergence ($\beta = 0$) are defined as functions of the parameter β:

$$d_\beta(x|y) = \begin{cases} \frac{1}{\beta(\beta-1)} \sum_{f=1}^{F} \sum_{t=1}^{T} \left(x_{f,t}^{\beta} + (\beta-1) y_{f,t}^{\beta} - \beta x_{f,t} y_{f,t}^{\beta-1} \right) & \beta \epsilon (0,1) \cup (1,2] \\ \sum_{f=1}^{F} \sum_{t=1}^{T} x_{f,t} log \frac{x_{f,t}}{y_{f,t}} - x_{f,t} + y_{f,t} & \beta = 1 \\ \sum_{f=1}^{F} \sum_{t=1}^{T} \frac{x_{f,t}}{y_{f,t}} - log \frac{x_{f,t}}{y_{f,t}} - 1 & \beta = 0 \end{cases} \quad (4)$$

where $x = X_{n_\beta}$ and $y = X_P + X_H$.

This factorization only considers the β-divergence cost and cannot automatically determine whether a basis belongs to a percussive signal or a harmonic signal. To overcome this drawback, the factorization process is modified to include specific spectro-temporal features related to percussive and harmonic sounds. Therefore, our contribution to the analysis of monaural signals is the development of a percussive/harmonic separation using an unsupervised NMF approach. This NMF approach models the mixture signal using an objective function that considers the β-divergence cost and common spectro-temporal features from the percussive/harmonic signals. Unlike other methods [4-6,8] that have been developed for percussive/harmonic separation, our approach does not use any labelled data to train the NMF bases maintaining competitive SDR results.

2.2.2 Percussive cost

The percussive cost is used to model percussive sounds by assuming smoothness in frequency and sparseness in time. Percussive sounds are typically represented as broadband signals, which have an energy that is confined to short time intervals. We define spectral smoothness (*SSM*), which is associated with the matrix W_P here, in the same way as continuity is defined in [22]:

$$\text{SSM} = \frac{T}{R_p} \sum_{r_p=1}^{R_p} \frac{1}{{\sigma_{W_{P_{r_p}}}}^2} \sum_{f=2}^{F} \left(W_{P_{f-1,r_p}} - W_{P_{f,r_p}} \right)^2 \quad (5)$$

where a high cost is assigned to large changes in the frequency between the bases $W_{P_{f,r_p}}$ and $W_{P_{f-1,r_p}}$ in adjacent frequency bins. Normalization is used to make the global objective function independent of the signal norm, i.e., the bases W_P are normalized by σ_{W_P}. The value σ_{W_P} of each percussive component r_p is calculated as $\sigma_{W_{P_{r_p}}} = \sqrt{\frac{1}{F} \sum_{f=1}^{F} W_{P_{f,r_p}}^2}$. To ensure that each percussive or harmonic cost has the same weight in the global objective function, each cost is normalized. Taking *SSM* into account, it is normalized by a factor equal to $\frac{T}{R_p}$. Thus, we avoid scaling problems that can arise from the number of frames, the number of frequency bins or the number of components considered.

Another percussive restriction can be applied to the temporal distribution of activations. Temporal sparseness has been previously used in [22,23] as the L1 norm of the activation gains to penalize solutions with nonzero gains. The concept of temporal sparseness (*TSP*) is defined following [22] but is applied to the matrix H_P. That is, a high cost is assigned to nonzero gains assuming that percussive sounds can be represented as transients with energies that are concentrated over short time intervals.

$$\text{TSP} = \frac{F}{R_p} \sum_{r_p=1}^{R_p} \sum_{t=1}^{T} \left| \frac{H_{P_{r_p,t}}}{\sigma_{H_{P_{r_p}}}} \right| \quad (6)$$

Similarly to the W_P treatment, the activations H_P are normalized by σ_{H_P}. The value σ_{H_P} of each percussive component r_p is calculated as $\sigma_{H_{P_{r_p}}} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} H_{P_{r_p,t}}^2}$. As mentioned above, the TSP is normalized by a factor $\frac{F}{R_p}$ such that the percussive costs are equally weighted.

2.2.3 Harmonic cost

The harmonic cost is used to model harmonic sounds by assuming smoothness in time and sparseness in frequency. These sounds can be considered to be stable sounds that exhibit a slow variation in amplitude over time with most of their energy being concentrated at the spectral peaks. We define temporal smoothness (TSM), which is associated with the H_H matrix here, in the same way as temporal continuity is defined in [22]. This smoothness is used to assign a high cost to large changes in time between the gains $H_{H_{r_h,t-1}}$ and $H_{H_{r_h,t}}$ in adjacent frames.

$$\text{TSM} = \frac{F}{R_h} \sum_{r_h=1}^{R_h} \frac{1}{{\sigma_{H_{H_{r_h}}}}^2} \sum_{t=2}^{T} \left(H_{H_{r_h,t-1}} - H_{H_{r_h,t}} \right)^2 \quad (7)$$

Normalization is used to make the global objective function independent to the signal norm; Thus, the H_H gains are normalized by σ_{H_H}. The value σ_{H_H} of each harmonic component r_h is calculated as $\sigma_{H_{H_{r_h}}} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} H_{H_{r_h,t}}^2}$. The harmonic cost TSM is normalized by a factor $\frac{F}{R_h}$.

Spectral sparseness (SSP) is defined following [22] but is applied to the matrix W_H. SSP assigns a high cost to

nonzero bases assuming that the harmonic sounds can be represented as a set of overtones in the frequency.

$$\mathrm{SSP} = \frac{T}{R_h} \sum_{r_h=1}^{R_h} \sum_{f=1}^{F} \left| \frac{W_{H_{f,r_h}}}{\sigma_{W_{H_{r_h}}}} \right| \tag{8}$$

Similarly to the treatment for H_H, the bases W_H are normalized by σ_{W_H}. The value σ_{W_H} of each harmonic component r_h is calculated as $\sigma_{W_{H_{r_h}}} = \sqrt{\frac{1}{F}\sum_{f=1}^{F} W_{H_{f,r_h}}^2}$. The harmonic cost SSP is normalized by a factor $\frac{T}{R_h}$.

2.2.4 Global percussive/harmonic NMF algorithm

The global objective function D, including the β-divergence, percussive and harmonic costs, is formulated as follows:

$$\begin{aligned} D = {} & d_\beta\left(X_{n_\beta} | \left(X_P + X_H\right)\right) + K_{\mathrm{SSM}}\mathrm{SSM} + K_{\mathrm{TSP}}\mathrm{TSP} \\ & + K_{\mathrm{TSM}}\mathrm{TSM} + K_{\mathrm{SSP}}\mathrm{SSP} \end{aligned} \tag{9}$$

Preliminary results do not show significant differences from initializing K_{SSM} and K_{TSM} with different values; thus, these parameters are set equal to each other (i.e., $K_{\mathrm{SSM}} = K_{\mathrm{TSM}}$). The parameters K_{TSP} and K_{SSP} are treated in the same way (i.e., $K_{\mathrm{TSP}} = K_{\mathrm{SSP}}$) because a similar behaviour is observed for these parameters in the preliminary studies. To determine if both sparseness and smoothness affect the performance of the separation process, we define the parameter $K_{\mathrm{SP}} = K_{\mathrm{TSP}} = K_{\mathrm{SSP}}$ to represent the sparseness terms and the parameter $K_{\mathrm{SM}} = K_{\mathrm{TSM}} = K_{\mathrm{SSM}}$ to represent the smoothness terms.

Following [20], we use the so-called multiplicative update rules (see Equation 10), such that D is non-increasing while ensuring non-negativity of the bases and the gains. These rules can be implemented using a gradient descent algorithm with an appropriate choice of the step size and are estimated for each scalar parameter Z by expressing the partial derivatives of the objective function $\frac{\partial D}{\partial Z}$ as the division between two positive terms $\left[\frac{\partial D}{\partial Z}\right]^-$ and $\left[\frac{\partial D}{\partial Z}\right]^+$:

$$Z = Z \odot \frac{\left[\frac{\partial D}{\partial Z}\right]^-}{\left[\frac{\partial D}{\partial Z}\right]^+} \tag{10}$$

where $\odot$ is the element-wise product operator and the fraction is the element-wise division operator.

Substituting the percussive basis W_P and the percussive gain H_P into Equation 10, the multiplicative update rules of the percussive sounds are formulated in Equations 11 and 12. The equations of each term $\left[\frac{\partial d_\beta}{\partial W_P}\right]^\pm$, $\left[\frac{\partial \mathrm{SSM}}{\partial W_P}\right]^\pm$, $\left[\frac{\partial d_\beta}{\partial H_P}\right]^\pm$ and $\left[\frac{\partial \mathrm{TSP}}{\partial H_P}\right]^\pm$ can be found in the Appendix.

$$W_P = W_P \odot \frac{\left[\frac{\partial d_\beta}{\partial W_P}\right]^- + K_{\mathrm{SSM}}\left[\frac{\partial \mathrm{SSM}}{\partial W_P}\right]^-}{\left[\frac{\partial d_\beta}{\partial W_P}\right]^+ + K_{\mathrm{SSM}}\left[\frac{\partial \mathrm{SSM}}{\partial W_P}\right]^+} \tag{11}$$

$$H_P = H_P \odot \frac{\left[\frac{\partial d_\beta}{\partial H_P}\right]^- + K_{\mathrm{TSP}}\left[\frac{\partial \mathrm{TSP}}{\partial H_P}\right]^-}{\left[\frac{\partial d_\beta}{\partial H_P}\right]^+ + K_{\mathrm{TSP}}\left[\frac{\partial \mathrm{TSP}}{\partial H_P}\right]^+} \tag{12}$$

Substituting the harmonic basis W_H and the harmonic gain H_H into Equation 10, the multiplicative update rules of the harmonic sounds are formulated in Equations 13 and 14. The equations of each term $\left[\frac{\partial d_\beta}{\partial W_H}\right]^\pm$, $\left[\frac{\partial \mathrm{SSP}}{\partial W_H}\right]^\pm$, $\left[\frac{\partial d_\beta}{\partial H_H}\right]^\pm$ and $\left[\frac{\partial \mathrm{TSM}}{\partial H_H}\right]^\pm$ can be found in the Appendix.

$$W_H = W_H \odot \frac{\left[\frac{\partial d_\beta}{\partial W_H}\right]^- + K_{\mathrm{SSP}}\left[\frac{\partial \mathrm{SSP}}{\partial W_H}\right]^-}{\left[\frac{\partial d_\beta}{\partial W_H}\right]^+ + K_{\mathrm{SSP}}\left[\frac{\partial \mathrm{SSP}}{\partial W_H}\right]^+} \tag{13}$$

$$H_H = H_H \odot \frac{\left[\frac{\partial d_\beta}{\partial H_H}\right]^- + K_{\mathrm{TSM}}\left[\frac{\partial \mathrm{TSM}}{\partial H_H}\right]^-}{\left[\frac{\partial d_\beta}{\partial H_H}\right]^+ + K_{\mathrm{TSM}}\left[\frac{\partial \mathrm{TSM}}{\partial H_H}\right]^+} \tag{14}$$

By iteratively updating the matrices W_P, H_P, W_H, H_H using *maxIter* iterations, our scheme can automatically distinguish between the bases belonging to the percussive or harmonic sounds.

2.2.5 Signal reconstruction

The percussive signal $x_p(t)$ is synthesized by using the magnitude percussive spectrogram X_P (see Equation 3) computed as the product of the factorized bases W_P and the activations H_P. To ensure that the reconstruction process is conservative, a percussive mask M_P is generated using Wiener filtering [24]. The effect of a percussive mask is to scale every frequency bin with a ratio that explains how much the percussive source contributes in the mixed spectrogram. The phase information related to the percussive signal is computed by multiplying the percussive mask M_P by the complex spectrogram X_c related to the mixed signal $x(t)$. The harmonic mask M_H is taken into account to similarly compute the harmonic signal $x_h(t)$.

$$M_P = \frac{X_P^2}{X_P^2 + X_H^2} \tag{15}$$

$$M_H = \frac{X_H^2}{X_P^2 + X_H^2} \tag{16}$$

$$x_p(t) = \mathrm{IDFT}\left(M_P \cdot X_c\right) \tag{17}$$

$$x_h(t) = \mathrm{IDFT}\left(M_H \cdot X_c\right) \tag{18}$$

The developed percussive/harmonic sound separation is detailed in Algorithm 1.

Algorithm 1 The developed percussive/harmonic sound separation

1 Compute the normalized magnitude spectrogram $X_{n\beta}$ of the input signal $x(t)$
2 Initialize W_P, H_P, W_H, H_H with random nonnegative values
3 **for** iteration=1:*maxIter* **do**
4 Update W_P using the multiplicative update rule (see Equation 11)
5 Update W_H using the multiplicative update rule (see Equation 13)
6 Update H_P using the multiplicative update rule (see Equation 12)
7 Update H_H using the multiplicative update rule (see Equation 14)
8 **end for**
9 Reconstruction of the percussive signal $x_p(t)$ (see Equation 17)
10 Reconstruction of the harmonic signal $x_h(t)$ (see Equation 18)

3 Evaluation and comparison

3.1 Test data

The databases are composed of monaural real-world music excerpts. Each music excerpt contains percussive and pitched instruments but does not contain a vocal track. The development database E (Table 1) is taken from the Guitar Hero game [25,26] and is composed of five commercial excerpts, each lasting 30 s. The first test database T1 (Table 2), taken from the Guitar Hero game [25,26], is composed of 20 commercial excerpts, each lasting 30 s. A pseudo-random process, using the standard uniform distribution, is used to select a starting time followed by a 30-s excerpt. The second test database T2 (Table 3) is a public database taken from SiSEC 2010 [27] and consists of four professionally produced music recordings lasting between 14 and 24 s. All of the signals were converted from stereo to mono and sampled at 16 kHz with a 16-bit resolution.

In summary, the dataset is composed of three databases. The development database E (five excerpts) is used to optimize the parameters (β, K_{SM}, K_{SP}, R_p and R_h) of the developed method. Two test databases T1 (20 excerpts) and T2 (4 excerpts) are then used to evaluate the performance of the separation process. Note that the development database E is not a part of the test databases T1 and T2.

Table 1 Identifier, title and artist of the files of the development database E

Identifier	Title	Artist
*E*_01	Two minutes to midnight	Iron Maiden
*E*_02	Bullet with butterfly wings	Smashing Pumpkins
*E*_03	Gamma ray	Beck
*E*_04	Go your own way	Fleetwood Mac
*E*_05	Hotel California	Eagles

3.2 Experimental setup

The quality of the audio separation using different frame sizes N, time shifts J and *maxIter* iterations is evaluated in preliminary studies. The experimental results show approximately the same performance using $N > 1024$ samples and $J < 512$ samples; therefore, we use the values $N = 1024$ and $J = 512$ because these values provide the best trade-off between the performance and the computational cost. The convergence of the algorithm is empirically observed. In fact, in all the performed simulations the convergence is achieved after 100 iterations. For this reason, we choose $maxIter = 100$. The values R_p and R_h are initialized to 150 because we initially supposed that this number of components would be a representative number of percussive and harmonic spectral

Table 2 Identifier, title and artist of the files of the first test database T1

Identifier	Title	Artist
*T1*_01	In my place	Coldplay
*T1*_02	La bamba	Los Lobos
*T1*_03	Livin' on a prayer	Bon Jovi
*T1*_04	No one to depend on	Santana
*T1*_05	Ring of fire	Johnny Cash
*T1*_06	Rooftops	Lost prophets
*T1*_07	So lonely	The Police
*T1*_08	Song 2	Blur
*T1*_09	Sultans of swing	Dire Straits
*T1*_10	Under pressure	Queen
*T1*_11	Are you gonna go my way	Lenny Kravitz
*T1*_12	Feel the pain	Dinosaur Jr
*T1*_13	Hollywood nights	Bob Seger & The Silver Bullet Band
*T1*_14	Hurts so good	John Mellencamp
*T1*_15	Kick out the jams	MC5's Wayne Kramer
*T1*_16	Make it wit chu	Queens Of The Stone Age
*T1*_17	One way or another	Blondie
*T1*_18	Shiver	Coldplay
*T1*_19	Shout it out loud	Kiss
*T1*_20	Sympathy for the devil	The Rolling Stones

Table 3 Identifier, title and artist of the files of the second test database T2

Identifier	Title	Artist
T2_01	Roads	Bearlin
T2_02	The_ones_we_love	Another Dreamer
T2_03	Remember_the_name	Fort Minor
T2_04	Tour	Ultimate NZ Tour

patterns. However, these parameters will be later analyzed (subsection 3.4.1).

Sound source separation applications, based on NMF approaches, usually adopt logarithmic frequency discretization (e.g., uniformly spaced sub-bands on the equivalent rectangular bandwidth (ERB) scale [28]). As harmonic signals are organized in a chromatic scale and using this scale, the musical notes are defined with semitone resolution, the developed method uses a 1/4 semitone resolution. Therefore, the time-frequency representation is obtained by integrating the STFT bins corresponding to each 1/4 semitone interval. To obtain the separated signals, the frequency resolution of the percussive and harmonic masks defined in Equations 15 and 16 must be extended to the resolution of the STFT. Taking into account that each bin of the STFT belongs to a value in the 1/4 semitone resolution, each bin of the masks with the STFT resolution takes the value that belongs to in the 1/4 semitone resolution. Consequently, all bins belonging to the same 1/4 semitone have the same mask value. Percussive and harmonic masks with the resolution of the STFT are then obtained and the inverse transform can be computed following Equations 17 and 18.

3.3 Algorithms for comparison

We use three recent state-of-the-art percussive/harmonic sound separation methods to evaluate the developed method: HPSS [1], MFS [7] and NMPCF [8]. HPSS and MFS are implemented in this study, whereas the separation results from NMPCF have been provided directly by the authors.

3.4 Results

Three metrics are used to assess the performance of the developed method [29,30]: (1) the source-to-distortion ratio (SDR), which provides information on the overall quality of the separation process; (2) the source-to-interferences ratio (SIR), which is a measure of the presence of harmonic sounds in the percussive signal and vice versa; and (3) the source-to-artifacts ratio (SAR), which provides information on the artifacts in the separated signal from separation and/or resynthesis.

As previously mentioned, the output of each method is composed of two signals, a percussive signal $x_p(t)$ (i.e., harmonic sounds have been attenuated or removed) and a harmonic signal $x_h(t)$ (i.e., percussive sounds have been attenuated or removed). In the database T1, the percussive average (Perc) is computed using the mean of all of the separated percussive signals. The harmonic average (Harm) is computed using the mean of all of the separated harmonic signals. In the same way, the overall average (Overall) is computed using the mean of all of the separated percussive and harmonic signals.

3.4.1 *Parameters optimization*

Figure 2 shows the optimization of the parameters β, K_{SP} and K_{SM} that are used to analyze the effect of the β-divergence cost and the weight of the smoothness and sparseness constraints in the development database E. Standard NMF (smoothness and sparseness constraints are disabled, i.e., $K_{\mathrm{SM}} = K_{\mathrm{SP}} = 0$) achieves a SDR value approximately equal to 3 dB. This fact suggests that standard NMF does not properly separate percussive and harmonic sounds because each separated signal, using standard NMF, is composed of percussive and harmonic sounds but its energy is approximately half of the energy of the mixed input signal. The results show that the maximum SDR is obtained using $\beta = 1.5$ because the SDR is maximized at the best trade-off between high and low energy changes in the frequency, unlike the Euclidean distance (which corresponds to $\beta = 2.0$ and is more sensitive to high energy changes) or the Itakura-Saito divergence (which corresponds to $\beta = 0$ and is more sensitive to low energy changes). This optimization produces a significant improvement of approximately 2.7 dB over the standard NMF showing that a percussive separated signal is composed of sounds that exhibit percussive features in which harmonic sounds have been attenuated, and the harmonic separated signal is composed of sounds that exhibit harmonic features where percussive sounds have been attenuated. As a consequence, the factorized spectrograms exhibit time-frequency energy distributions such as can be found in real-world percussive or harmonic sounds. In order to obtain the maximum SDR, the percussive and harmonic costs are found to be equally significant ($K_{\mathrm{SM}} = K_{\mathrm{SP}} = 0.1$) but must be sufficiently small compared to the β-divergence cost to reconstruct the signal correctly. As a result of this fact, the developed method fails for $K_{\mathrm{SM}} >> 1$ or $K_{\mathrm{SP}} >> 1$ because NMF does not prioritize signal reconstruction under these conditions.

Figure 2 has shown the optimization of β, K_{SP} and K_{SM} using a fixed dictionary size ($R_p = R_h = 150$). In order to analyze the dependence of the parameters of the developed method, Figures 3, 4 and 5 are shown.

Figure 3 shows the optimization of the parameters K_{SP} and K_{SM} using $\beta = 1.5$ and different number of percussive R_p and harmonic R_h components in order to analyze the effect of different dictionary sizes. Results

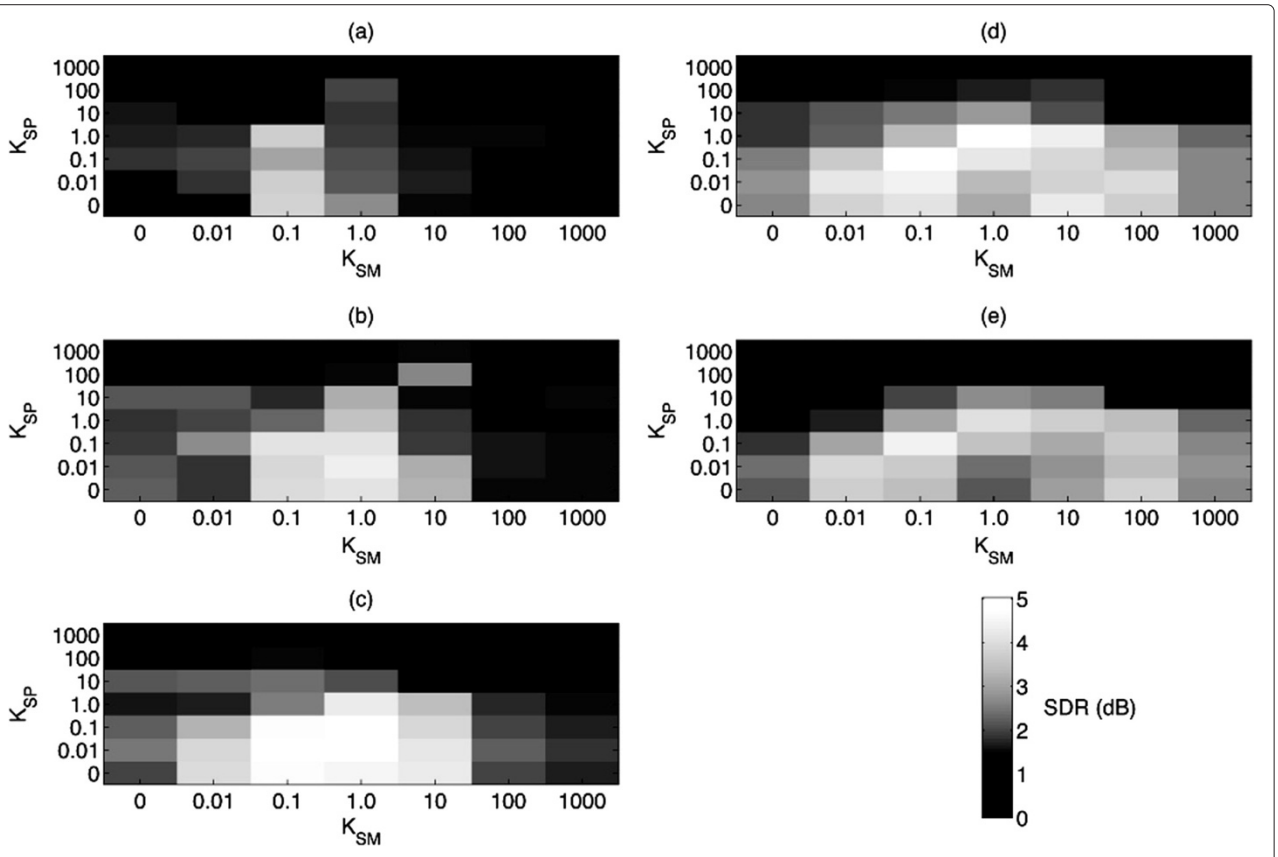

Figure 2 Optimization of β, K_{SP} and K_{SM} parameters using the development database E. (a) $\beta = 0$, **(b)** $\beta = 0.5$, **(c)** $\beta = 1.0$, **(d)** $\beta = 1.5$ and **(e)** $\beta = 2.0$. The number of percussive R_p and harmonic R_h spectral patterns is equal to 150. It can be observed the effect of the β-divergence cost in the separation process when varying the contribution of the smoothness and sparseness constraints. The goal of this analysis was to find the optimum value of parameters β, K_{SP} and K_{SM} looking for the value of the combination of them that maximizes the SDR measure for the development database E. The maximum SDR value is obtained using $\beta = 1.5$. This fact may be due to the best trade-off between high and low energy changes in frequency.

show the optimal values of the parameters K_{SP} and K_{SM} are obtained using $K_{SP} = K_{SM} = 0.1$ for all the dictionary sizes evaluated. This fact suggests that the dictionary size could not affect the optimal values of K_{SP} and K_{SM} because these values provide the maximum SDR for each dictionary size evaluated.

Figure 4 shows the optimization of the parameter β using the optimal $K_{SP} = K_{SM} = 0.1$ and different number of percussive and harmonic components in order to analyze the effect of a different dictionary sizes. For each value β, results show that SDR performance exhibits the similar behaviour independently of the dictionary size, increasing from $\beta = 0$ to $\beta = 1.5$ and decreasing from $\beta = 1.5$ to $\beta = 2$. As occurred in Figure 2, the value β that maximizes SDR is achieved using $\beta = 1.5$, but the differences, compared to the other results using different dictionary sizes, are not significant.

Figure 5 shows the optimization of the smoothness K_{TSM}-K_{SSM} and sparseness K_{TSP}-K_{SSP} parameters. Figure 5a shows that although different values of K_{TSM} and K_{SSM} have been used, the maximum SDR performance is obtained using $K_{TSM} = K_{SSM} = 0.1$ as occurred when these parameters were initialized with the same values (see Figure 2d). This fact suggests that the smoothness parameters K_{TSM} and K_{SSM} could be initialized using equal values $K_{SM} = K_{TSM} = K_{SSM}$ since they do not show significant differences from initializing them with different values in order to obtain the best SDR performance. A similar behaviour can be observed from the sparseness K_{TSP}-K_{SSP} parameters (see Figure 5b) which obtain the maximum SDR performance using a sparseness parameter equal to each other $K_{SP} = K_{TSP} = K_{SSP} = 0.1$.

Figure 6 shows how SDR results can be improved by analyzing a more accurate range of the parameters K_{SM} and K_{SP} around of the previous optimum value 0.1. The maximum SDR is obtained using a higher value of smoothness constraints than sparseness constraints, specifically $K_{SM} = 0.2$ and $K_{SP} = 0.1$, evaluating the development database E. This fact can be observed by

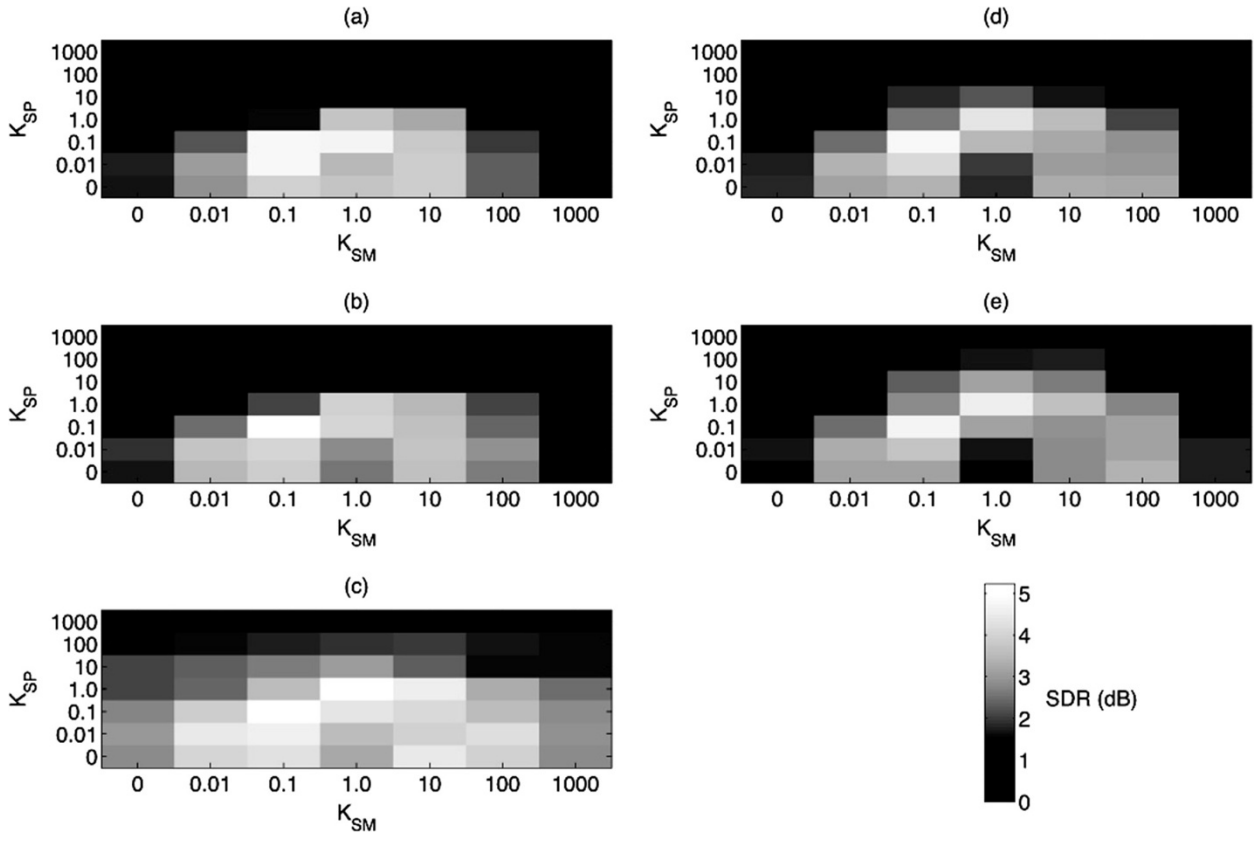

Figure 3 Optimization of parameters K_{SP} and K_{SM}. Using $\beta = 1.5$ and different dictionary sizes composed of the same number of percussive and harmonic components (**(a)** 50, **(b)** 100, **(c)** 150, **(d)** 250, and **(e)** 500 components) using the development database E.

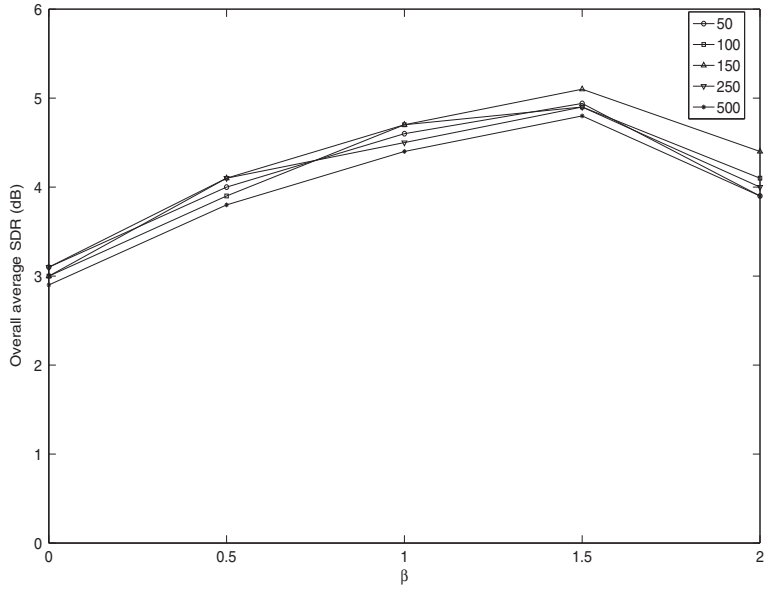

Figure 4 Optimization of the parameter β. Using the optimal $K_{SP} = K_{SM} = 0.1$ and different dictionary sizes composed of the same number of percussive and harmonic components (50, 100, 150, 250 and 500 components) in the development database E.

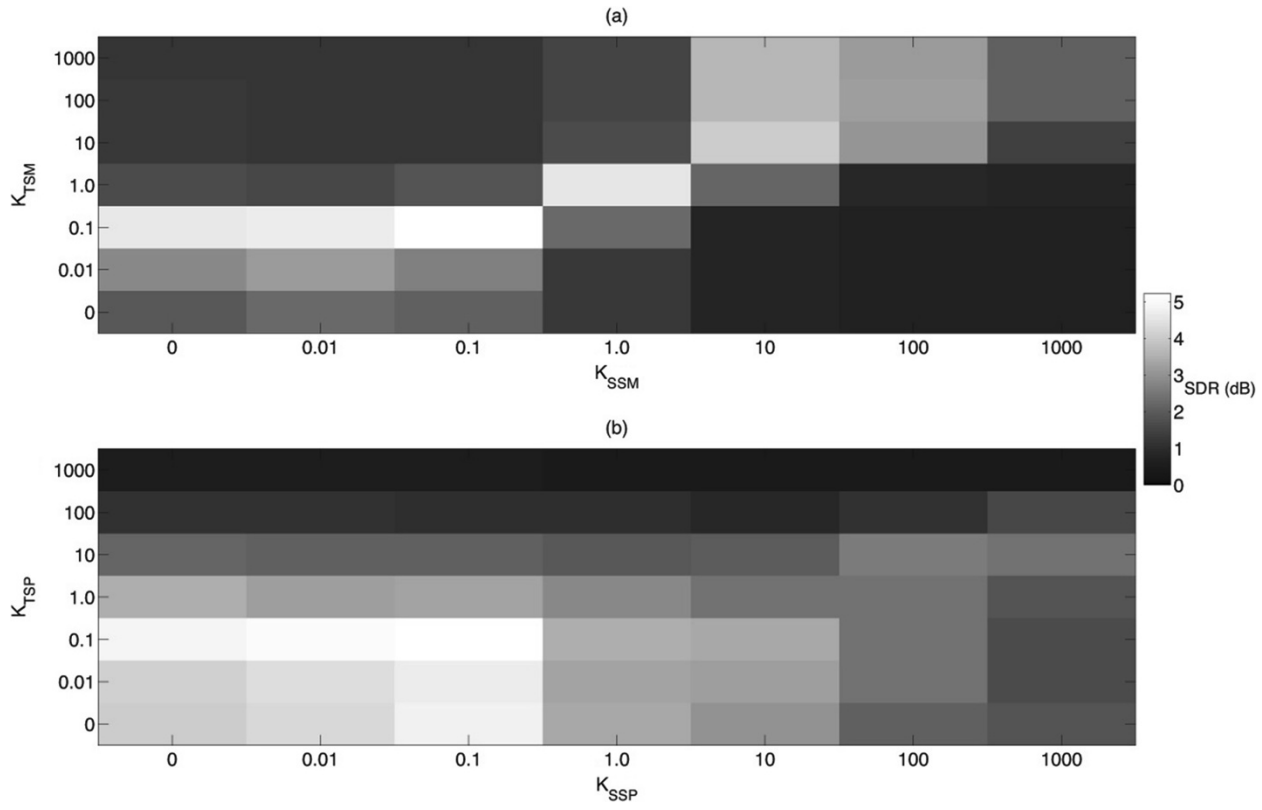

Figure 5 Optimization of (a) the smoothness and (b) sparseness parameters. (a) Smoothness parameters K_{TSM} and K_{SSM} using the values $\beta = 1.5$, $K_{\mathrm{TSP}} = K_{\mathrm{SSP}} = 0.1$ and $R_p = R_h = 150$ components in the development database E; **(b)** sparseness parameters K_{TSP} and K_{SSP} using the values $\beta = 1.5$, $K_{\mathrm{TSM}} = K_{\mathrm{SSM}} = 0.1$ and $R_p = R_h = 150$ components in the development database E.

analyzing the diagonals in Figure 6. Thus, using a K_{SM} value that is twice the K_{SP} value (i.e., see the lower diagonals from the main diagonal from left to right) improves the SDR results and slows the decrease in the SDR. However, when K_{SP} is twice K_{SM} (i.e., see the upper diagonals from the main diagonal from left to right), the SDR results are generally worse, and the SDR decreases more rapidly.

Once β, K_{SM} and K_{SP} are optimized ($\beta = 1.5, K_{\mathrm{SM}} = 0.2$ and $K_{\mathrm{SP}} = 0.1$), the number of the percussive R_p and harmonic R_h components is optimized (see Figure 7). A slight better separation is obtained when it is used $R_p = 250$ and $R_h = 500$. It seems that a wider variety of harmonic spectral patterns could improve SDR results; however, the improvement, about $0.3dB$, compared to $R_p = R_h = 150$ is not significant. In fact, a higher number of components, i.e., $R_p > 250$ and $R_h > 500$, reduces the SDR performance (in a similar way as occurred in Figure 3). This reduction of SDR may be explained by using a large dictionary size (usually $R_p + R_h$ is chosen to be smaller than F or T, so that $F(R_p + R_h) + (R_p + R_h)T << FT$ in a NMF framework [14,21]).

3.4.2 Performance evaluation

Figures 8, 9 and 10 show SDR, SIR and SAR results evaluating the database T1 for the proposed method and the three aforementioned state-of-the-art percussive/harmonic sound separation methods. Each box represents 20 data points, one for each excerpt of the test database: each data point in the blue box represents the average value of the percussive separation results; each data point in the red box represents the average value of the harmonic separation results; and each data point in the black box represents the average value of the overall separation results (the average value considering the percussive and harmonic separation results). The lower and upper lines of each box show the 25th and 75th percentiles for the database. The line in the middle of each box represents the mean value of the dataset. The lines extending above and below each box show the extent of the rest of the samples, excluding outliers. Outliers are defined as points that are over 1.5 times the interquartile range from the sample median, which are shown as crosses.

Figure 8 shows that the developed method obtains, on average, the best quality performance in terms of the percussive, harmonic and overall SDR for the separation

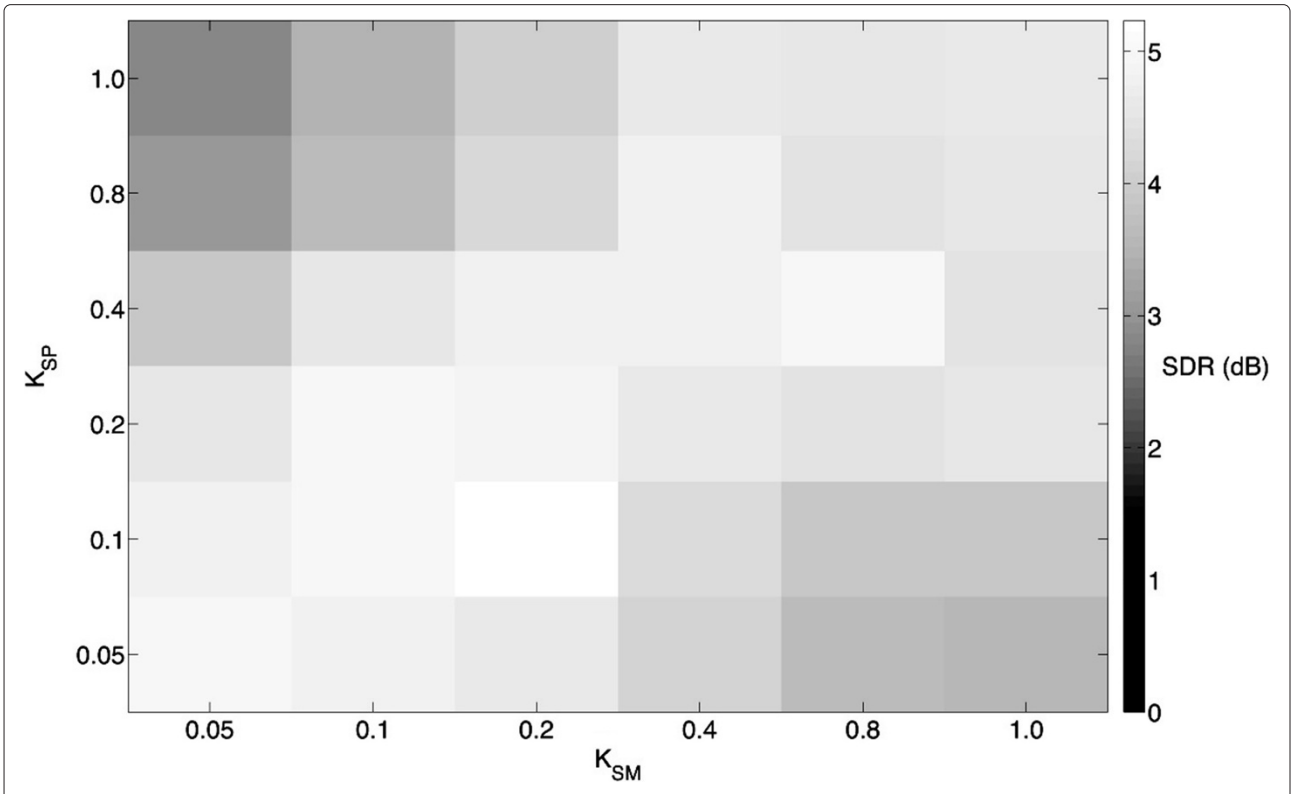

Figure 6 SDR average results. Obtained by using $\beta = 1.5$ and varying the smoothness term K_{SM} and the sparseness term K_{SP} for the development database E. The number of percussive R_p and harmonic R_h spectral patterns is equal to 150.

process relative to the three state-of-the-art separation methods. MFS and HPSS can still be considered to be competitive methods, unlike NMPCF (this method will be analyzed later). The proposed method significantly outperforms HPSS and NMPCF in percussive/harmonic SDR results but not compared to MFS which is confirmed by a one-sided paired t-test (see Table 4). A possible strength of the developed method, which is not exhibited by the other methods, seems to be its robustness in evaluating different databases (databases T1 and T2). This robustness is in the fact of including smoothness and sparseness constraints into the factorization process because the main difference between the develop method and HPSS and MFS is the use of sparseness constraints to achieve time-frequency energy distributions as can be found in real-world percussive or harmonic sounds. In both databases, the developed method produces nearly identical overall SDR results, 6.3 dB, independent of the database evaluated. Moreover, the lower line of each box of the developed method is above the mean obtained from using HPSS and NMPCF. Thus, more reliable SDR results are obtained in comparison with those methods when evaluating different types of sounds used in different music genres.

Figure 9 shows that HPSS produces the best overall SIR results but the SIR performance of the developed method is nearly identical to that obtained using HPSS. It can be confirmed using a one-sided paired t test that indicates HPSS does not significantly outperform SIR results compared to the developed method neither in percussive nor harmonic sounds (see Table 4). However, Table 4 shows that the developed method improves significantly SIR results compared to MFS and NMPCF for both percussive and harmonic sounds. Both HPSS and the proposed one: (i) enable most of the harmonic content to be removed while maintaining the quality of the percussive signal and vice versa and (ii) capture polyphonic richness. The developed method produces better percussive quality for the SIR than MFS because it uses information to model percussive sounds that is not used by MFS. That is, the developed method models percussive sounds using smoothness in the frequency and sparseness in the time, whereas MFS models percussive sounds using only smoothness in the frequency (by removing outliers in a temporal slice).

As previously mentioned, HPSS produces the best overall SIR on average. However, these results are obtained at the expense of introducing more artifacts and lead to greater overall distortion. This fact can be observed in Figure 10 in which the worst percussive and harmonic SAR results are obtained by HPSS. Considering SAR

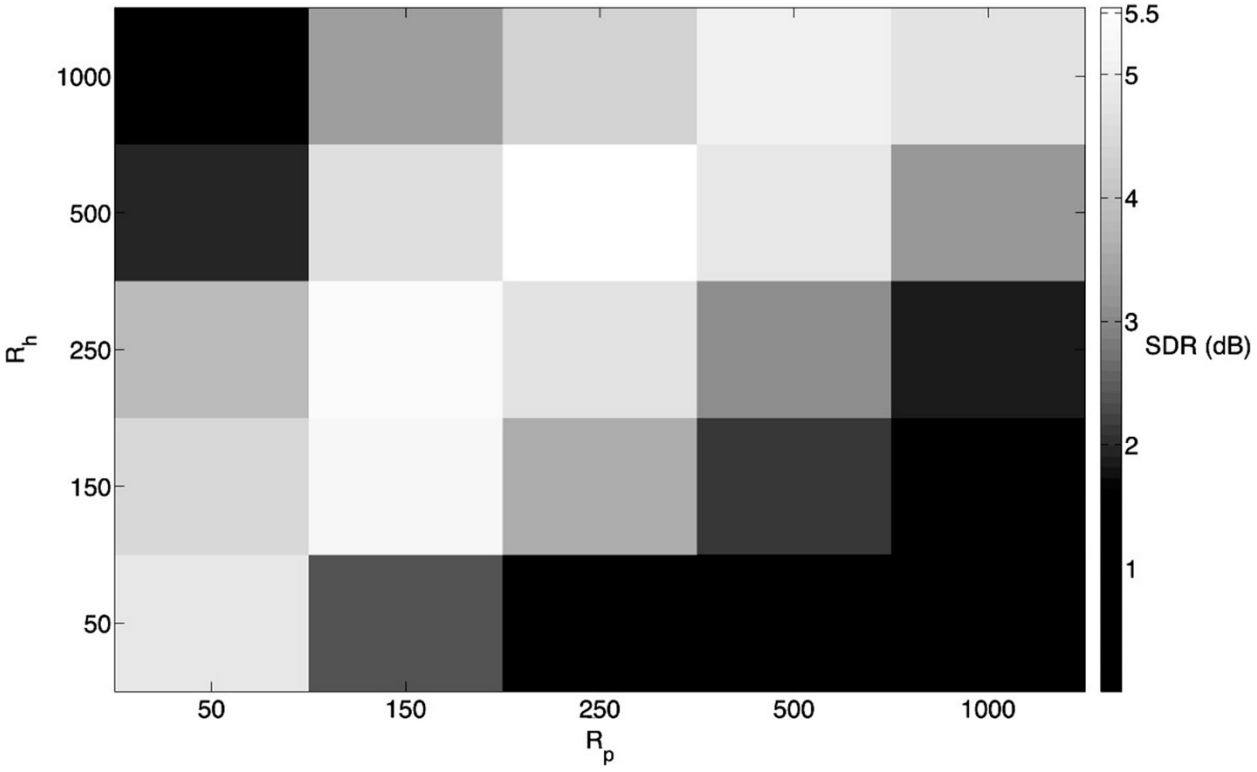

Figure 7 SDR average results ($\beta = 1.5$, $K_{SM} = 0.2$ and $K_{SP} = 0.1$). As a function of the number of percussive/harmonic components applied to the development database E.

results and using a one-sided paired *t* test (see Table 4), the developed method significantly outperforms HPSS and NMPCF in percussive sounds and the three state-of-the-art methods in harmonic sounds. The developed method also offers the advantage of producing the best SAR results (excluding NMPCF, which will be discussed in the next paragraph) because the artifacts in the reconstruction signal are minimized.

For the case of NMPCF, SDR and SIR results exhibit the worst separation performance, and therefore, this method

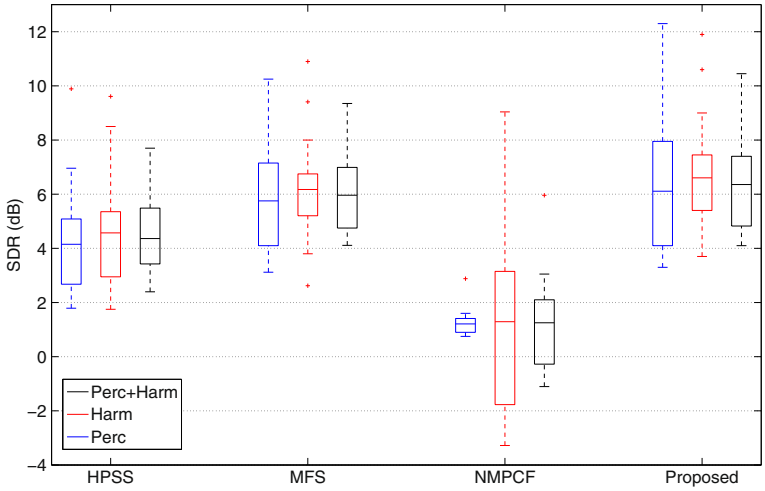

Figure 8 Comparative SDR results. Evaluating database T1 for the proposed method and the three aforementioned state-of-the-art percussive/harmonic sound separation methods.

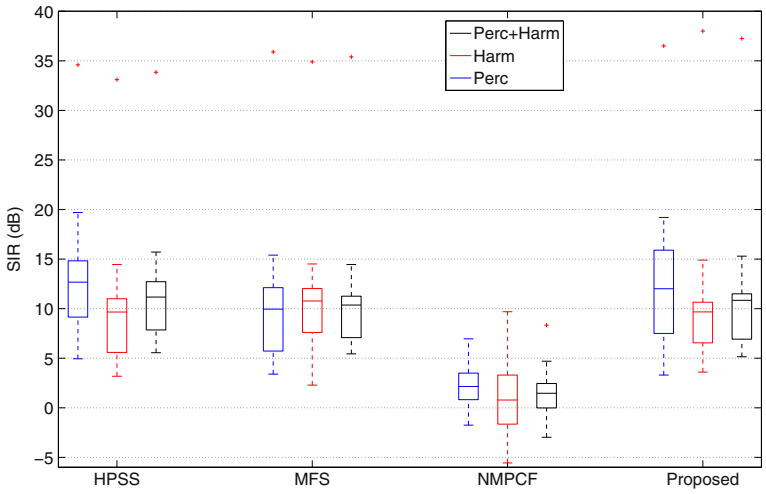

Figure 9 Comparative SIR results. Evaluating the database T1 for the proposed method and the three aforementioned state-of-the-art percussive/harmonic sound separation methods.

always ranks last. The poor performance of NMPCF may be attributed to its high dependence on the drum-only matrix used in the decomposition process. This drum-only matrix is obtained training with drum sounds, but these trained drum features may be sufficiently different from the percussive features evaluated in the test database T1. The harmonic signal, provided by NMPCF, is composed of most of the original harmonic and percussive sounds. It causes a high harmonic SAR because the proportion of artifacts is too small compared to the proportion of the target (harmonic) and the interference (percussive) sounds. However, the percussive signal, provided by NMPCF, is composed of a residual part of the original percussive and harmonic sounds. It causes a low percussive SAR because the proportion of artifacts is too high compared to the proportion of the target (percussive) and the interference (harmonic) sounds.

To illustrate some of the strengths and weaknesses of the developed method, Table 5 shows the SDR, SIR and SAR results for each individual music excerpt from database T2. Results show that the developed method produces the best average percussive, harmonic and overall SDR and SAR results. The main strength of the developed method is its high separation performance in evaluating purely harmonic or percussive sounds (e.g., SAR (dB) bass or drums). However, HPSS obtains the better SIR results on average with more than 1 dB above the proposed method and MFS. For all the methods evaluated, the

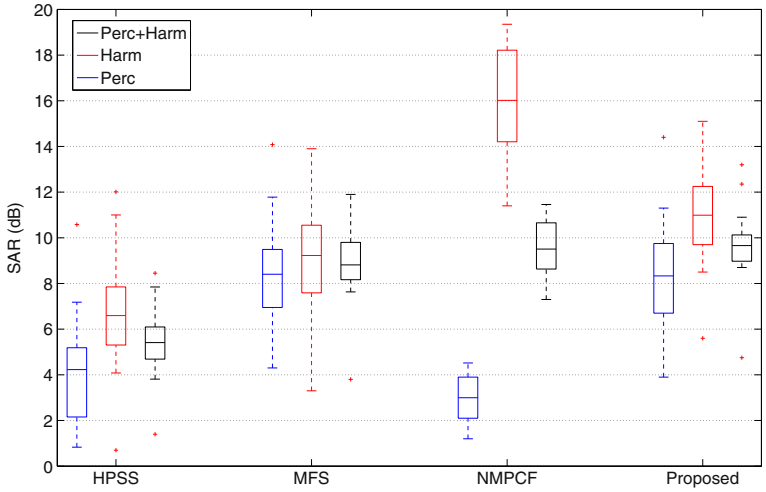

Figure 10 Comparative SAR results. Evaluating the database T1 for the proposed method and the three aforementioned state-of-the-art percussive/harmonic sound separation methods.

Table 4 Analysis of the statistical significance of the percussive/harmonic SDR-SIR-SAR results

	Percussive			Harmonic		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
HPSS	$< 10^{-8}$	0.2	$< 10^{-11}$	$< 10^{-9}$	0.9	$< 10^{-14}$
MFS	0.2	$< 10^{-3}$	0.8	0.1	$< 4 \cdot 10^{-2}$	$< 10^{-6}$
NMPCF	$< 10^{-7}$	$< 10^{-4}$	$< 10^{-7}$	$< 10^{-6}$	$< 10^{-4}$	$< 10^{-5}$

Comparing the developed method with the three state-of-the-art separation methods using a one-sided paired *t* test in the database T1. Each cell show the parameter *p* that represents the probability of setting a statistically significant result. Considering a confidence interval of 95%, small values of $p<0.05$ indicate that there exists statistical significance of the results evaluated.

main weaknesses of harmonic/percussive separation are the following: (i) It is not effective in separating harmonic onsets, that is, the transients of harmonic sounds played by harmonic instruments (e.g., the initial milliseconds of a note played by guitar) because they exhibit a percussive behaviour. In these cases, harmonic onsets are not separated in the harmonic signal as can be seen in Figure 11. The reason is because a harmonic onset exhibits spectro-temporal features that have been modelled as percussive sounds in the factorization process; (ii) it cannot effectively separate audio effects, e.g., a synthesizer can generate a harmonic sound that exhibits spectro-temporal features (e.g., the vibrato effect shows non-smoothness in time) which are not modelled in the factorization process.

To illustrate the separation performance of the developed method, audio examples (from the T1 and T2 databases) have been uploaded to a web page. Each audio example (mixed track, separated-percussive track and separated-harmonic track) has been evaluated using HPSS, MFS and the developed method. The web page can be found at https://dl.dropboxusercontent.com/u/22448214/PercHarmFeb2014/index.html.

4 Conclusions

This paper presents an unsupervised learning process for separating percussive and harmonic sounds from monaural instrumental music. Our formulation is based on a modified NMF approach that automatically distinguishes between percussive and harmonic bases by integrating spectro-temporal features, such as anisotropic smoothness or time-frequency sparseness, into the factorization

Table 5 Percussive, harmonic and overall SDR, SIR and SAR results for each excerpt of the database T2

	HPSS			MFS			Proposed		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Percussive separation									
Identifier									
T2_01	2.6	13.2	1.1	−0.2	−1.5	8.4	4.0	6.5	5.7
T2_02	2.4	10.2	3.4	3.1	8.0	4.9	5.2	8.3	7.5
T2_03	2.6	6.9	4.0	2.5	2.1	12.3	2.8	2.6	11.1
T2_04	5.5	11.5	6.5	6.2	9.6	8.0	7.5	10.3	10.3
Average	3.2	*10.5*	3.8	2.9	4.6	8.4	*4.9*	7.0	*8.7*
Harmonic separation									
Identifier									
T2_01	9.8	13.8	11.9	7.1	13.8	11.5	11.0	14.8	13.9
T2_02	4.8	6.3	9.8	5.5	16.2	11.6	7.5	9.3	12.1
T2_03	4.8	8.7	6.3	4.6	11.0	8.0	5.0	9.1	8.6
T2_04	5.6	11.5	6.7	6.2	9.3	8.7	7.5	10.6	10.5
Average	6.3	10.1	8.7	5.9	*12.6*	10.0	*7.8*	11.0	*11.3*
Overall separation									
Identifier									
T2_01	6.2	13.5	6.5	3.5	6.2	10.0	7.5	10.7	9.8
T2_02	3.6	8.3	6.6	4.3	12.1	8.3	6.4	8.8	9.8
T2_03	3.7	7.8	5.2	3.6	6.6	10.2	3.9	5.9	9.9
T2_04	5.6	11.5	6.6	6.2	9.5	8.4	7.5	10.5	10.4
Average	4.8	*10.3*	6.2	4.4	8.6	9.2	*6.3*	9.0	*10.0*

Each row titled Average shows three italic values. Each italic value represents the best percussive, harmonic or overall SDR, SIR or SAR result comparing all methods evaluated in this table.

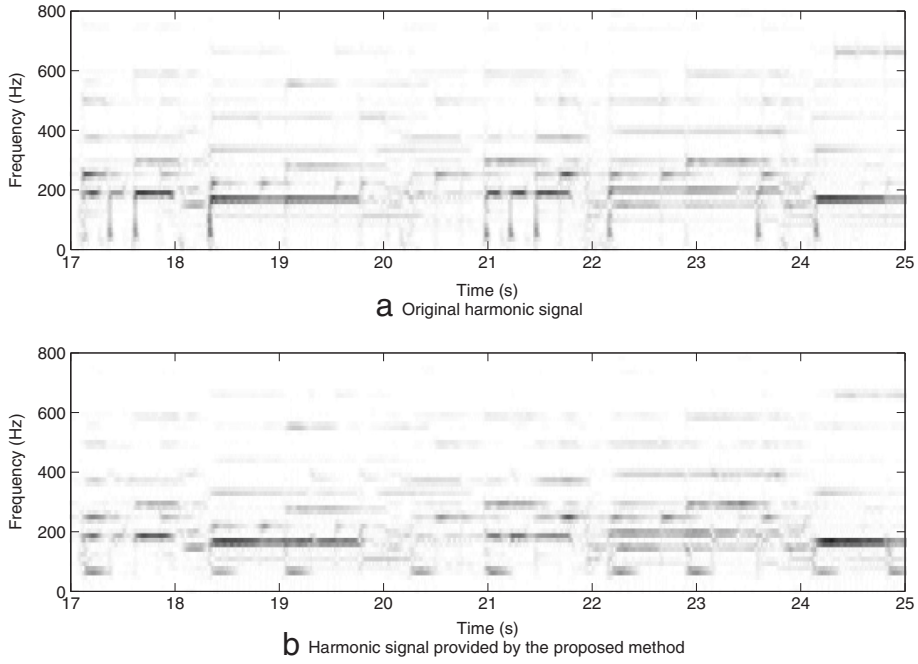

Figure 11 A harmonic excerpt, played by guitar instrument, extracted from the file $T2\_02$ in Table 3. It can be observed that the harmonic onsets (vertical lines at the beginning of some notes) in the original harmonic signal **(a)** do not appear in the harmonic signal **(b)** output of the proposed method.

process. The developed method exhibits the following advantages: (i) prior knowledge of the number of instruments playing the music excerpt is not required, and (ii) neither prior information about the musical instruments nor a supervised training are required to classify the bases.

Different experiments are performed to optimize the parameters of the developed method. The results show that (i) the value $\beta = 1.5$ provides the maximum SDR since this measure has been computed at the best trade-off between high and low energy changes in the frequency; (ii) The maximum SDR is achieved using a higher value of smoothness constraints compared to sparseness constraints evaluating the databases T1 and T2 and (iii) a higher number of components, i.e., $R_p > 250$ and $R_h > 500$, reduces the SDR performance.

The analysis of the dependence of the parameters of the developed method shows that (i) the dictionary size could not affect the optimal values of K_{SP} and K_{SM} because they provide the maximum SDR for each dictionary size evaluated; (ii) SDR performance obtains the maximum SDR using $\beta = 1.5$ independently of the dictionary size and (iii) the smoothness K_{TSM}-K_{SSM} and sparseness K_{TSP}-K_{SSP} parameters could be initialized using equal values $K_{\mathrm{SM}} = K_{\mathrm{TSM}} = K_{\mathrm{SSM}}$ and $K_{\mathrm{SP}} = K_{\mathrm{TSP}} = K_{\mathrm{SSP}}$ since they do not show significant differences from initializing them with different values in order to obtain the maximum SDR performance.

Evaluating the database T1 shows that the developed method obtains the best quality performance in terms of the percussive, harmonic and overall SDR for the separation process in relation to the three state-of-the-art separation methods. The proposed method significantly outperforms the other methods taking into account most of the percussive/harmonic SDR, SIR or SAR results which is confirmed by a one-sided paired t test. A significant strength of the developed method is its robustness in evaluating different databases.

Evaluating the database T2 illustrates some of the strengths and weaknesses of the developed method. An interesting strength shown by the developed method is its successful separation performance in evaluating purely harmonic or percussive sounds. However, harmonic onsets and audio effects are not successfully separated because their spectro-temporal features have not been modelled in the factorization process.

Future work will focus on three topics. First, we will try to improve the quality of the separated signals by defining a new spectral distance that integrates novel spectro-temporal features of the percussive and harmonic sounds. Second, a novel constraint based on the vibrato effect will be investigated to extend this method to singing-voice signals. Finally, in order to improve the singing-voice signals, a set of novel percussive constraints will be analyzed to distinguish between percussive music instruments and unvoiced vocal sounds.

Appendix

Detailed terms of the multiplicative update rules of percussive sounds

Here, each of the terms $\left[\frac{\partial d_\beta}{\partial W_P}\right]^{\pm}$, $\left[\frac{\partial \mathrm{SSM}}{\partial W_P}\right]^{\pm}$, $\left[\frac{\partial d_\beta}{\partial H_P}\right]^{\pm}$ and $\left[\frac{\partial \mathrm{TSP}}{\partial H_P}\right]^{\pm}$ belonging to the percussive multiplicative update rules are detailed:

$$\left[\frac{\partial d_\beta}{\partial W_P}\right]^{-} = \left[(W_P H_P + W_H H_H)^{\beta-2} \odot X_{n_\beta}\right] H_P^T \quad (19)$$

$$\left[\frac{\partial d_\beta}{\partial W_P}\right]^{+} = \left[(W_P H_P + W_H H_H)^{\beta-1}\right] H_P^T \quad (20)$$

$$\left[\frac{\partial \mathrm{SSM}}{\partial W_P}\right]^{-}_{f,r_p} = 2F\left[\frac{W_{P_{f-1,r_p}} + W_{P_{f+1,r_p}}}{\sum_{j=1}^{F} W^2_{P_{j,r_p}}}\right] + \frac{2FW_{P_{f,r_p}} \sum_{j=2}^{F}\left(W_{P_{j,r_p}} - W_{P_{j-1,r_p}}\right)^2}{\left(\sum_{j=1}^{F} W^2_{P_{j,r_p}}\right)^2} \quad (21)$$

$$\left[\frac{\partial \mathrm{SSM}}{\partial W_P}\right]^{+}_{f,r_p} = \frac{4FW_{P_{f,r_p}}}{\sum_{j=1}^{F} W^2_{P_{j,r_p}}} \quad (22)$$

$$\left[\frac{\partial d_\beta}{\partial H_P}\right]^{-} = W_P^T\left[(W_P H_P + W_H H_H)^{\beta-2} \odot X_{n_\beta}\right] \quad (23)$$

$$\left[\frac{\partial d_\beta}{\partial H_P}\right]^{+} = W_P^T\left[(W_P H_P + W_H H_H)^{\beta-1}\right] \quad (24)$$

$$\left[\frac{\partial \mathrm{TSP}}{\partial H_P}\right]^{-}_{r_p,t} = \sqrt{T}\frac{H_{P_{r_p,t}} \sum_{i=1}^{T} H_{P_{r_p,i}}}{\left(\sum_{i=1}^{T} H^2_{P_{r_p,i}}\right)^{\frac{3}{2}}} \quad (25)$$

$$\left[\frac{\partial \mathrm{TSP}}{\partial H_P}\right]^{+}_{r_p,t} = \frac{1}{\sqrt{\frac{1}{T}\sum_{i=1}^{T} H^2_{P_{r_p,i}}}} \quad (26)$$

where ^{T} denotes the transpose matrix operator. The terms $\left[\frac{\partial \mathrm{SSM}}{\partial W_P}\right]^{\pm}$ and $\left[\frac{\partial \mathrm{TSP}}{\partial H_P}\right]^{\pm}$ are defined using [22], as adapted to the matrix W_P and H_P.

Detailed terms of the multiplicative update rules of harmonic sounds

Here, each of the terms $\left[\frac{\partial d_\beta}{\partial W_H}\right]^{\pm}$, $\left[\frac{\partial \mathrm{SSP}}{\partial W_H}\right]^{\pm}$, $\left[\frac{\partial d_\beta}{\partial H_H}\right]^{\pm}$ and $\left[\frac{\partial \mathrm{TSM}}{\partial H_H}\right]^{\pm}$ belonging to the harmonic multiplicative update rules are detailed:

$$\left[\frac{\partial d_\beta}{\partial W_H}\right]^{-} = \left[(W_P H_P + W_H H_H)^{\beta-2} \odot X_n\right] H_H^T \quad (27)$$

$$\left[\frac{\partial d_\beta}{\partial W_H}\right]^{+} = \left[(W_P H_P + W_H H_H)^{\beta-1}\right] H_H^T \quad (28)$$

$$\left[\frac{\partial \mathrm{SSP}}{\partial W_H}\right]^{-}_{f,r_h} = \sqrt{F}\frac{W_{H_{f,r_h}} \sum_{j=1}^{F} W_{H_{j,r_h}}}{\left(\sum_{j=1}^{F} W^2_{H_{j,r_h}}\right)^{\frac{3}{2}}} \quad (29)$$

$$\left[\frac{\partial \mathrm{SSP}}{\partial W_H}\right]^{+}_{f,r_h} = \frac{1}{\sqrt{\frac{1}{F}\sum_{j=1}^{F} W^2_{H_{j,r_h}}}} \quad (30)$$

$$\left[\frac{\partial d_\beta}{\partial H_H}\right]^{-} = W_H^T\left[(W_P H_P + W_H H_H)^{\beta-2} \odot X_n\right] \quad (31)$$

$$\left[\frac{\partial d_\beta}{\partial H_H}\right]^{+} = W_H^T\left[(W_P H_P + W_H H_H)^{\beta-1}\right] \quad (32)$$

$$\left[\frac{\partial \mathrm{TSM}}{\partial H_H}\right]^{-}_{r_h,t} = 2T\left[\frac{H_{H_{r_h,t-1}} + H_{H_{r_h,t+1}}}{\sum_{i=1}^{T} H^2_{H_{r_h,i}}}\right] + \frac{2TH_{H_{r_h,t}} \sum_{i=2}^{T}\left(H_{H_{r_h,i}} - H_{H_{r_h,i-1}}\right)^2}{\left(\sum_{i=1}^{T} H^2_{H_{r_h,i}}\right)^2} \quad (33)$$

$$\left[\frac{\partial \mathrm{TSM}}{\partial H_H}\right]^{+}_{r_h,t} = \frac{4TH_{H_{r_h,t}}}{\sum_{i=1}^{T} H^2_{H_{r_h,i}}} \quad (34)$$

where ^{T} denotes the transpose matrix operator. The terms $\left[\frac{\partial \mathrm{SSP}}{\partial W_H}\right]^{\pm}$ and $\left[\frac{\partial \mathrm{TSM}}{\partial H_H}\right]^{\pm}$ are defined using [22], as adapted to the matrix W_H and H_H.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the Andalusian Business, Science and Innovation Council under project P2010- TIC-6762 and (FEDER) the Spanish Ministry of Economy and Competitiveness under Project TEC2012-38142-C04-03.

Author details

[1]Telecommunication Engineering Department, University of Jaen, Linares, Jaen, Spain. [2]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain.

Received: 23 February 2014 Accepted: 12 June 2014
Published: 11 July 2014

References

1. N Ono, K Miyamoto, J Le Roux, H Kameoka, 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* (Lausanne Switzerland, August 2008), pp. 25–29
2. N Ono, K Miyamoto, H Kameoka, S Sagayama, A real-time equalizer of harmonic and percussive components in music signals, in *Proceedings of the Ninth International Conference on Music Information Retrieval (ISMIR)* (Philadelphia, Pennsylvania USA, September 14–18 2008), pp. 139–144
3. L Daudet, Review on techniques for the extraction of transients in musical signals, in *Proceedings of the Third International Conference on Computer Music Modeling and Retrieval* (Pisa, Italy, September 26–28 2005), pp. 219–232

4. M Helen, T Virtanen, Separation of drums from polyphonic music using non-negative matrix factorisation and support vector machine, in *Proceedings of the European Signal Processing Conference* (Anatalya, Turkey, September 4–8 2005)
5. O Gillet, G Richard, Transcription and separation of drum signals from polyphonic music. IEEE Trans. Audio Speech Lang. Process. **3**(16), 529540 (2008)
6. A Ozerov, E Vincent, F Bimbot, A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. Audio Speech Lang. Process. **20**(4), 11181133 (2012)
7. D Fitzgerald, Harmonic/percussive separation using median filtering, in *Proceedings of DAFX* (Graz, Austria, September 6–10 2010)
8. J Yoo, M Kim, K Kang, 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)* (Dallas, Texas, USA, March 14–19 2010)
9. R Jain, R Kasturi, B Schunck, *Machine Vision*. (McGraw-Hill, New York, 1995)
10. H Tachibana, H Kameoka, S Sagayama, Comparative evaluations of various harmonic/percussive sound separation algorithms based on anisotropic continuity of spectrogram, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Tokyo, Japan, March 25–30 2012)
11. F Canadas-Quesada, N Ruiz-Reyes, P Vera-Candeas, J Carabias, S Maldonado, A multiple-F0 estimation approach based on Gaussian spectral modelling for polyphonic music transcription. J. New Music Res. **39**(1), 93–107 (2010)
12. Y Ueda, Y Uchiyama, T Nishimoto, N Ono, S Sagayama, *HMM-based approach for automatic chord detection using refined acoustic features*, (Dallas, Texas, USA, March 14–19 2010)
13. D Zhiyao, B Pardo, A state space model for online polyphonic audio-score alignment, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Prague, Czech Republic, May 22–27 2011)
14. D Lee, S Seung, Learning the parts of objects by nonnegative matrix factorization. Nature. **401**(21), 788–791 (1999)
15. P Hoyer, Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004)
16. V Monga, M Mhcak, Robust and secure image Hashing via non-negative matrix factorizations. IEEE Trans. Inf. Forensics Secur. **2**(3), 376–390 (2007)
17. I Kotsia, S Zafeiriou, I Pitas, A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. IEEE Trans. Inf. Forensics Secur. **2**(3), 588–595 (2007)
18. P Smaragdis, J Brown, Non-negative matrix factorization for polyphonic music transcription, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, New York, USA, October 19–22 2003)
19. J Paulus, T Virtanen, Drum transcription with non-negative spectrogram factorisation, in *Proceedings of the European Signal Processing Conference* (Antalya, Turkey, September 4–8 2005)
20. D Lee, H Seung, Algorithms for non-negative matrix factorization, in *Advances in NIPS*, (2000), pp. 556–562
21. C Févotte, N Bertin, JL Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Comput. **21**(3), 793830 (2009)
22. T Virtanen, Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**(3), 1066–1074 (2007)
23. J Eggert, E Korner, Sparse coding and NMF, in *Proceedings of the International Joint Conference on Neural Networks (IJCNN4)* (Budapest, Hungary, 25–29 July 2004), pp. 2529–2533
24. J Parras-Moral, F Canadas-Quesada, P Vera-Candeas, N Ruiz-Reyes, Audio restoration of solo guitar excerpts using a excitation-filter instrument model, in *Stockholm Music Acoustics Conference jointly with Sound And Music Computing Conference* (Stockholm, Sweden, 30 July)
25. Activision, Guitar hero World Tour. http://en.wikipedia.org/wiki/Guitar_Hero_World_Tour. Accessed 09/06/2014
26. Activision, Guitar hero 5. http://en.wikipedia.org/wiki/Guitar_hero_5. Accessed 09/06/2014
27. S Araki, A Ozerov, V Gowreesunker, H Sawada, F Theis, G Nolte, D Lutter, N Duong, The 2010 signal separation evaluation campaign (SiSEC2010): audio source separation, in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)* (Saint-Malo France, September 2010), pp. 114–122
28. E Vincent, Musical source separation using time-frequency source priors. IEEE Trans. Audio Speech Lang. Process. **14**(1), 91–98 (2006)
29. E Vincent, C Févotte, R Gribonval, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
30. C Févotte, R Gribonval, E Vincent, BSS_EVAL toolbox user guide - Revision, 2.0, Technical Report 1706, IRISA (April 2005)

doi:10.1186/s13636-014-0026-5
Cite this article as: Canadas-Quesada *et al.*: **Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:26.