A Method to Separate Musical Percussive Sounds using Chroma Spectral Flatness

F.J. Cañadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, A. Muñoz-Montoro, F.J. Bris-Peñalver

Telecommunication Engineering Department, University of Jaen
Scientific and Technological Campus, Cinturon Sur s/n, 23700
Linares, Jaen, Spain

Email: fcanadas@ujaen.es, pvera@ujaen.es, nicolas@ujaen.es, antoniojmmontoro@gmail.com,fjbris@gmail.com

***Abstract*—This paper presents an unsupervised Non-Negative Matrix Factorization (NMF) approach to extract percussive sounds from monaural music signals. Due to unconstrained NMF cannot discriminate between percussive, harmonic or singing-voice components in the decomposition process, we propose a novel method to extract percussive sounds based on the anisotropic smoothness of percussive chroma. Thus, percussive sounds can be discriminate because chroma from percussive sounds clearly draws lines along the chroma. Under a NMF framework, a time-domain signal related to a component is labelled as percussive is the energy distribution of its chroma is approximately flat. This proposal does not require information about the number of active sound sources neither prior knowledge about the instruments nor supervised training to classify the bases. Real-world audio mixtures composed of Harmonic/Percussive and Harmonic/Percussive/Singing-voice sounds were evaluated. Experimental results showed that the proposal was effective compared to state-of-the-art methods. An interesting advantage of the proposal is that it can remove most of the singing-voice components from the extracted percussive signals.**

Keywords–Non-negative matrix factorization; Sound source separation; monaural; percussive; chroma; spectral flatness; distortion;

I. INTRODUCTION

The extraction of percussive sounds from monaural audio mixtures has received much attention over the last decade. Percussive sounds, e.g., snare drum, are impulsive and are typically smooth in frequency. Harmonic sounds, e.g, bass or piano, are quasi-stationary and are typically smooth in time. Therefore, percussive sounds have a structure that is vertically smooth in frequency, whereas harmonic sounds have a structure that is horizontally smooth in time. However, singing-voice sounds are not smooth in frequency because most of them are composed of spectral peaks located at integer multiples of the fundamental frequency and are not smooth in time due to pitch fluctuations (e.g., vibrato effect) as can be seen in Figure 1. Specifically, Figure1 shows that percussive sounds draw vertical lines whereas harmonic sounds draw horizontal lines. Singing-voice sounds draw fluctuated lines over the time. A method capable of separating percussive sounds from audio can be used to facilitate a wide range of Music Information Retrieval (MIR) applications. Some of these include onset detection, beat tracking, rhythm pattern recognition, remixing and for audio to score alignment.

Several approaches have exploited the concept of anisotropic smoothness which is related to the difference in the directions of continuity between the spectrograms of harmonic and percussive sounds. Ono et al. [1] [2] separate harmonic and percussive sounds by exploiting the anisotropy of harmonic and percussive sounds in a Maximum A Posteriori (MAP) framework. Fitzgerald's system [3] extracts percussive sounds using the anisotropy smoothness by means of a median filtering. In this manner, the harmonics are considered to be outliers in a temporal slice that contains a mixture of percussive and pitched instruments. In [4], percussive extraction is performed using non-negative matrix partial co-factorization. Thus, the shared basis vectors in this co-factorization are associated with the percussive features, which are used to extract drum-related components from audio.

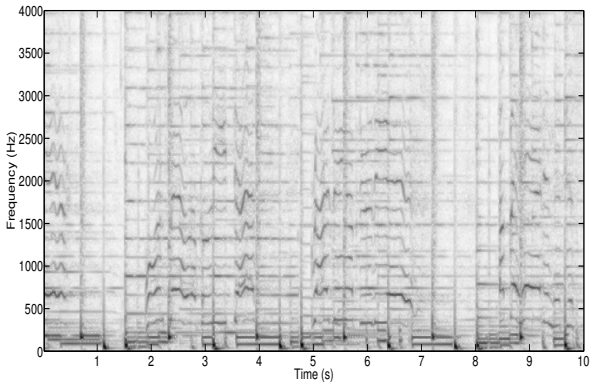

Figure 1. Spectrogram of an audio excerpt composed of percussive, harmonic and singing-voice sounds.

Recently, a measure [5] based on a segmental spectral flatness is used to distinguish between harmonic and percussive signals. However, evaluation was performed using mixtures composed of one harmonic source and one percussive source not providing experimental results with commercial real-world excerpts. Becker et al. [6] propose an extension that supports spectral continuity and a new temporal continuity constraint using temporal flatness. Canadas et al. [7] propose an unsupervised learning process based on a modified Non-Negative Matrix Factorization (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 process.

In this paper, we propose an intuitive, novel and fast method to separate percussive sounds from monaural music. Using the concept of anisotropic smoothness, in a similar way

 ISBN: 978-1-61208-487-9

as the spectrogram of a percussive sound draws a line along the frequency direction, the chroma of a percussive sound draws a line along the 12 distinct semitones. A time-domain signal related to a component decomposed by NMF can be labelled as percussive if the energy distribution of its chroma is approximately flat.

The remainder of the paper is organized as follows. Section II introduces NMF and its application to sound source separation briefly. Section III describes the proposed method. Experimental results and performance analysis are shown in Section IV. Conclusions and future work are reported in Section V.

II. BACKGROUND

NMF [8] is a technique for multivariate data analysis which aims to obtain a parts-based representation of objects, by imposing non-negative constraints. Given a matrix $\mathbf{X}$ of dimensions $F \times T$ with non-negative entries, it is possible to model it as linear combinations of K elementary non-negative spectra. Therefore, NMF is the problem of finding a factorization:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{WH} \tag{1}$$

where $\hat{\mathbf{X}}$ is the estimated matrix, $\mathbf{W} \in \mathbb{R}^{F \times K}$ is the matrix whose columns are the bases, spectral patterns or components. $\mathbf{H} \in \mathbb{R}^{K \times T}$ is a matrix of component gains or activations for all frames. K is usually chosen such that $FK + KT << FT$, hence reducing the data dimension. In typical audio applications, the matrix $\mathbf{X}$ is chosen as a time-frequency representation (e.g., magnitude or power spectrogram), $f = 1, ..., F$ denoting the frequency bin and $t = 1, ..., T$ the time frame.

In the case of magnitude spectra, the parameters are restricted to be non-negative, then, a common way to compute the factorization in Eq. (1) is generally obtained by minimizing a cost function defined as

$$D(\mathbf{X}|\hat{\mathbf{X}}) = \sum_{f=1}^{F} \sum_{t=1}^{T} d(X_{ft}|\hat{X}_{ft}) \tag{2}$$

where $d(a|b)$ is a function of two scalar variables, d is typically non-negative and takes value zero if and only if $a = b$. In this work, the generalized Kullback-Leibler divergence has been used since it is the most frequently used cost function in sound source separation [9] and our preliminary experiments showed that the generalized Kullback-Leibler divergence obtained better separation performance compared to Euclidean distance and the Itakura Saito divergence [10].

An iterative algorithm based on multiplicative update rules is proposed in [8] to obtain the model parameters that minimize the cost function. Under these rules, the generalized Kullback-Leibler divergence $D_{KL}(\mathbf{X}|\hat{\mathbf{X}})$ is non-increasing at each iteration and it is ensured the non negativity of the bases and the gains [8].

$$D_{KL}(\mathbf{X}|\hat{\mathbf{X}}) = \sum_{f} \sum_{t} \mathbf{X}_{ft} \log \frac{\mathbf{X}_{ft}}{\hat{\mathbf{X}}_{ft}} - \mathbf{X}_{ft} + \hat{\mathbf{X}}_{ft} \tag{3}$$

The update rules can be defined as follows,

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \frac{\mathbf{X}}{(\mathbf{WH})}}{\mathbf{W}^T \mathbf{1}_{F,T}} \tag{4}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{\mathbf{X}}{(\mathbf{WH})} \mathbf{H}^T}{\mathbf{1}_{F,T} \mathbf{H}^T} \tag{5}$$

where $\mathbf{W}$ and $\mathbf{H}$ are initialized as random positive matrices, $\mathbf{1}_{F,T}$ represents a matrix of all-one composed of F rows and T columns, ^{T} is the transpose operator, $\odot$ represents the Hadamard (element-wise) multiplication and the division is also element-wise.

III. PROPOSED METHOD

An intuitive, novel and fast method to extract percussive sounds in music recordings is proposed. It is composed of three stages (NMF, Chroma and Spectral flatness) shown in Figure 2.

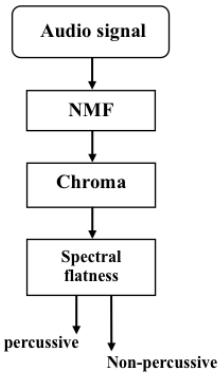

Figure 2. Block diagram of the proposed percussive separation method

The main idea of the proposal is based on the concept of anisotropic smoothness. Instead of using the anisotropic smoothness with the spectrogram data [2] [3], we use the anisotropic smoothness with the chroma data. To obtain the chroma data, a time-frequency representation is used in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. As a result, the chroma representation reports the intensity of each of the 12 distinct musical chroma of the octave at each time frame [11]. Just like a spectrogram of percussive sounds draw lines along the frequency direction, Figure 3 shows that chroma of a percussive sound also draws lines along the 12 distinct semitones because percussive sounds are characterized by smoothness in frequency. Therefore, our aim is to classify what components from NMF are percussive using the information provided by the energy distribution of the chroma.

In a first stage, the magnitude of the Short-Time Fourier Transform (STFT) $\mathbf{X}$ of a music signal $x(t)$, with a complex spectrogram $\mathbf{X_c}$ composed of T frames and F frequency bins, is calculated (details are shown in section IV-B). Using the generalized Kullback-Leibler divergence as previously mentioned, an unconstrained NMF is applied to the input spectrogram $\mathbf{X}$ using the update rules Eq. (4)-(5) obtaining a set of K bases or components.

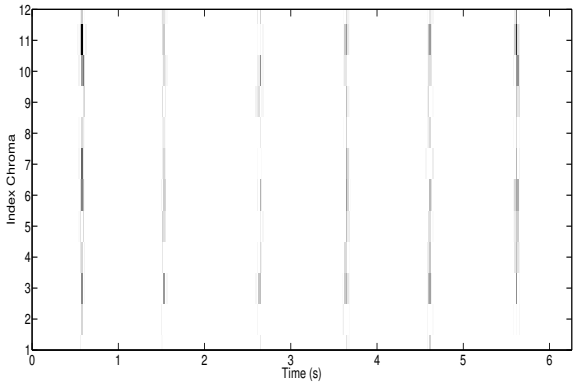

Figure 3. Chroma of a percussive time-domain reconstructed component obtained from NMF. There exists six drum sounds along the time.

In a second stage, each time-domain signal $x_i(t)$ from each component i^{th} is synthesized by the inverse overlap-add STFT of the product of the basis W_i and the activations H_i related to the component i^{th} and using the phase spectrogram of the input signal $x(t)$. Next, the chroma matrix [11] of the signal $x_i(t)$ is calculated generating a sequence of 12 frequency bins and T short-time frames and finally a chroma vector C_i is calculated summing all frames.

Initially, we applied the measure spectral flatness [12] directly on the bases W_i obtained in the NMF decomposition but results showed that it does not work because this measure is very sensitive to small values [5]. In order to overcome this problem, we propose to compute the chroma spectral flatness SF_i because each chroma bin has lower probability of having a small value. The measure chroma spectral flatness SF_i is computed using the measure spectral flatness [12] on the chroma vector C_i instead of basis vector W_i to avoid the high dependency of the spectral flatness related to the small values. $SF_i = 0$ implies a perfect harmonic sound while a $SF_i = 1$ implies a perfect percussive sound.

$$SF_i = \frac{(\prod_{k=1}^{12} C_i(k))^{\frac{1}{12}}}{\frac{1}{12}\sum_{k=1}^{12} C_i(k)} \in [0, 1] \quad (6)$$

Therefore, the extracted percussive signal $x_p(t)$ is the sum of all the signals $x_i(t)$ whose spectral flatness SF_i is higher than a threshold U. In this manner, $x_p(t)$ is composed of all signals $x_i(t)$ whose energy distribution of its chroma is approximately flat.

IV. Evaluation

A. Data set, metrics and State-of-the-art methods

Two data sets T1 and T2, composed of the same nine monaural real-world music excerpts, taken from the Guitar Hero game [13] [14], have been generated to evaluate the proposed method as can be seen in Table I. To perform an objective evaluation, each music excerpt from database T1 was created mixing the original percussive and harmonic instrumental tracks without using any singing-voice track. However, each music excerpt from database T2 was created mixing the same percussive and harmonic instrumental tracks of the database T1 and the original singing-voice track. Each excerpt has a duration about 30 seconds. All of the signals were converted from stereo to mono and sampled at 16 kHz.

TABLE I. IDENTIFIER, TITLE AND ARTIST OF THE FILES OF THE DATABASES T1 AND T2

IDENTIFIER	TITLE	ARTIST
$M1$	Hollywood Nights	Bob Seger & The Silver Bullet Band
$M2$	Hotel California	Eagles
$M3$	Hurts So Good	John Mellencamp
$M4$	La Bamba	Los Lobos
$M5$	Make It Wit Chu	Queens Of The Stone Age
$M6$	Ring of Fire	Johnny Cash
$M7$	Rooftops	Lost prophets
$M8$	Sultans of Swing	Dire Straits
$M9$	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) [15] [16] 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 non-percussive sounds in the percussive 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 [15].

We compare the separation performance of the proposed method with two reference percussive and harmonic sound separation methods. The first one is the method HPSS [2] and the second one is the method MFS [3].

B. Parameters

The STFT of each mixture has been calculated using half-overlapping Hamming window of $L = 1024$ samples, corresponding to a duration of 64 milliseconds at a sampling rate of 16KHz [1] [2] [7].

A random initialization of the matrices W and H was used and the convergence of the NMF decomposition was empirically observed which was achieved after 100 iterations. Due to NMF is not guaranteed to find a global minimum, the performance of the proposed method depends on the initial values of NMF [9] leads to different results. For this reason, we have repeated three times for each excerpt and the results in the paper are averaged values.

Highlight that the best separation performance will be obtained using an optimization process which is data dependent because NMF is a blind decomposition method and it is not based on the physics of the problem. As a result, the separated signals from NMF are not independent, nor uncorrelated, it generates many false positives and/or mixing of percussive/non-percussive sounds. In our preliminary results, we have evaluated different numbers of components $K = 10, 20, 30, 40$ and we selected the value $K = 10$ because it obtained the best separation performance. It seems that a small number of components improves the percussive separation because the subspace of percussive sounds is of lower rank compared to harmonic or singing-voice rank [17] [18].

C. Optimization

Figure 4 and Figure 5 show the optimization of the threshold U in the database T1 and T2. A lower value

 ISBN: 978-1-61208-487-9

of $U < 0.5$ captures higher percentage of non-percussive sounds that implies a reduction of the percussive SIR. A value around $U = 0.6$ allows a promising discrimination between percussive and non-percussive sounds. The optimum value of the threshold $U_{o1} = 0.6$ in the database T1 and $U_{o2} = 0.7$ in the database T2 have been selected because reach the best percussive and non-percussive SDR and a high percussive SIR associated with the highest non-percussive SIR. This situation reports that the proposed system extracts a high percentage of percussive sounds avoiding the extraction of non-percussive sounds and viceversa. It can be seen that higher values of $U > 0.7$ lose a high amount of percussive sounds that causes a drastically reduction of the non-percussive SIR. Comparing Figure 4 and Figure 5, an optimum threshold $U_{o2} > U_{o1}$ must be set because the system needs to be more strict to separate sounds from singing-voice active in the database T2.

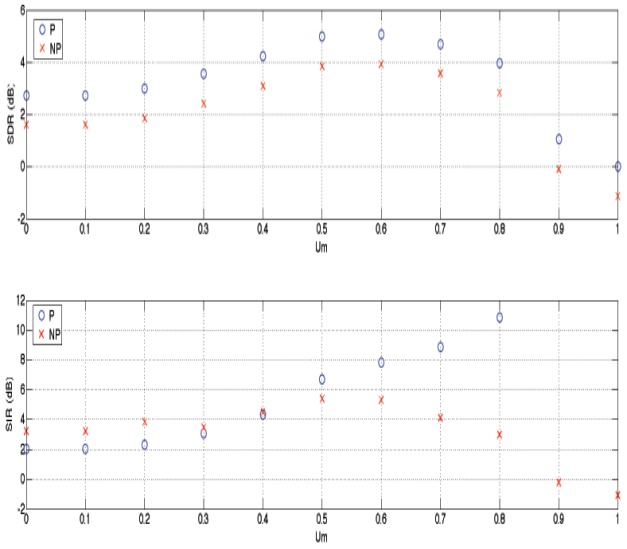

Figure 4. Average SDR-SIR obtained in function of the threshold U_m for the database T1. The legend P is related to percussive results and the legend NP is related to non-percussive results.

The optimum thresholds U_{o2} and U_{o1} will be used in the next section in order to evaluate the performance of the proposed system.

D. Results

Figure 6 and Figure 7 show SDR, SIR and SAR results evaluating the database T1 and T2 for the proposed method and the two state-of-the-art methods. Each box represents nine data points, one for each excerpt of the test database. Each method evaluated shows three boxes in figures. The left box represents the average value of the percussive separation results. The center box represents the average value of the non-percussive (harmonic sounds in the database T1 and harmonic+singing-voice sounds in database T2) separation results. The right box represents the overall average value considering the percussive and non-percussive 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 median 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

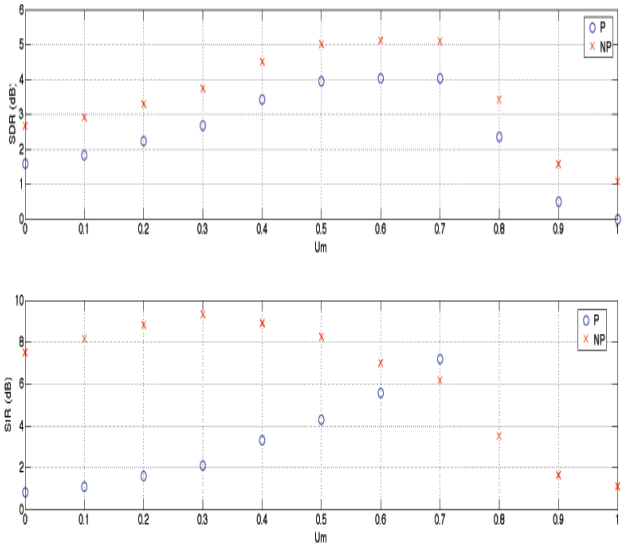

Figure 5. Average SDR-SIR obtained in function of the threshold U_m for the database T2. The legend P is related to percussive results and the legend NP is related to non-percussive results.

1.5 times the interquartile range from the sample median, which are shown as crosses.

Figure 6 displays the SDR separation performance for the database T1. It shows that MFS and the proposed method obtain the best percussive separation performance but HPSS can be considered as competitive method. Moreover, MFS achieves the best SDR taking into account non-percussive and overall separation followed by the proposed method. Taking into account SIR results, HPSS produces the best percussive SIR and MFS provides the best non-percussive SIR. However, the non-percussive SIR of the proposed method is the worst of them. This performance is because not all the bases decomposed by NMF and labelled as percussive are purely percussive bases (ideally, each component represents parts of a single sound source). It implies that some of the non-percussive sounds are also synthesized as percussive sounds by the proposed method. Therefore, the proposed method depends on the randomized initialization of the two matrices W and H in the NMF decomposition. Taking into account SAR results, HPSS achieves a high percussive SIR at the expense of introducing more artifacts, which it can be observed by the worst percussive and harmonic SAR. Nevertheless, the proposed method provides the best percussive and non-percussive SAR results because the artifacts in the reconstruction signal are minimized.

Figure 7 displays the separation performance for the database T2. Hereafter, all the comments are related to the comparison between Figure 6 and Figure 7. It can be observed that the addition of the singing-voice in the non-percussive sounds reduces about 1dB the overall SDR both MFS and HPSS but not in the proposed method. Specifically, the proposed method improves the overall SDR in about 1dB, obtaining approximately the same overall SDR that MFS. While HPSS and MFS reduces its percussive SIR about 5dB, the SIR reduction of the proposed method is only about 0.7dB. This performance indicates that a high amount of singing-voice sounds are active in the separated percussive signals

 ISBN: 978-1-61208-487-9

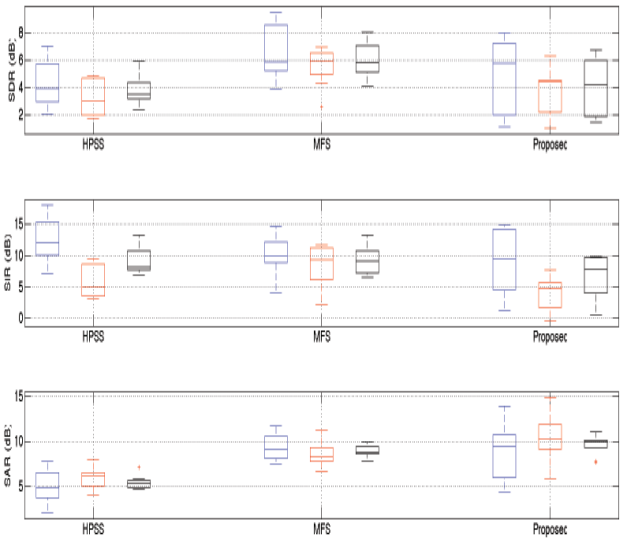

Figure 6. Separation performance in SDR, SIR and SAR results evaluating the database T1. The left box represents the average value of the percussive separation results. The center box represents the average value of the non-percussive separation results. The right box represents the overall average value considering the percussive and non-percussive separation results.

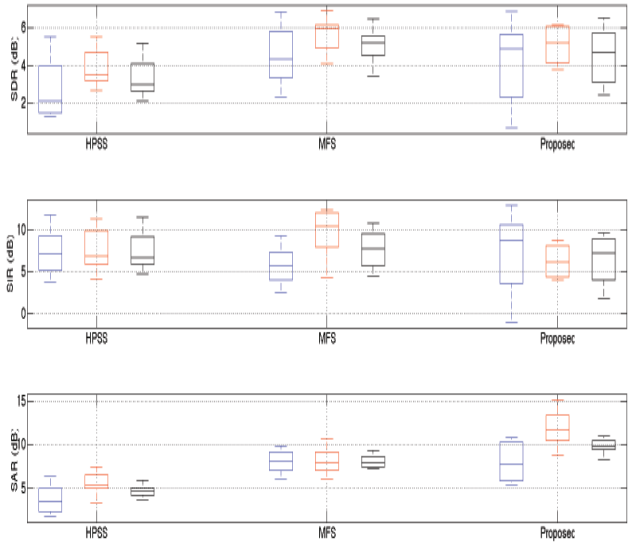

Figure 7. Separation performance in SDR, SIR and SAR results evaluating the database T2. The left box represents the average value of the percussive separation results. The center box represents the average value of the non-percussive separation results. The right box represents the overall average value considering the percussive and non-percussive separation results.

from HPSS and MFS. Moreover, the overall SIR in HPSS and MFS is reduced more than twice compared with the proposed method. As occurred in Figure 6, the proposed method achieves the best SAR results minimizing the artifacts in the reconstruction signal. Therefore, results report a strength of the proposed method, which is not exhibited by the other compared methods, that is the capability to successfully remove the singing voice sounds in the separated percussive signal. This capability implies that the proposed method provides the best tradeoff between the quality of the separated percussive signal and the removal of the singing voice sounds. An example of the mentioned capability to remove the singing voice sounds in the separated percussive signal is shown in Figures 8-10. It can be clearly observed that in the spectral range [400Hz-1600Hz] that most singing-voice sounds, characterized by fluctuated frequencies over the time, have only been removed using the proposed method.

To illustrate the separation performance of the proposed method, some percussive audio examples have been uploaded to a web page [19].

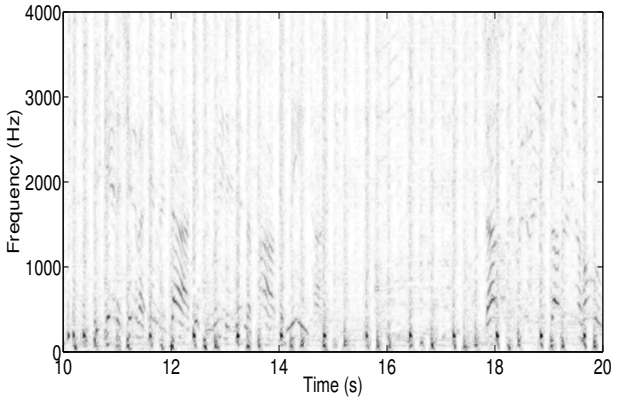

Figure 8. Percussive separation of the method HPSS evaluating the interval [10s-20s] of the file M8 of the database T2. Higher grey level represents higher energy of each frequency.

V. Conclusion

A novel, intuitive and fast method to separate percussive sounds from music, composed by percussive/harmonic instruments and singing-voice, is presented. Due to the fact that unconstrained NMF cannot discriminate between percussive, harmonic or singing-voice components in the decomposition process, we propose to extract percussive sounds based on the anisotropic smoothness of chroma. If the energy distribution of its chroma is approximately flat, then a time-domain signal related to a component decomposed by NMF can be labelled as percussive. This proposal does not require prior knowledge about the instruments nor supervised training to classify the bases.

Although the separation performance of the proposed method is competitive in evaluating mixtures of percussive and harmonic instruments, its performance depends on the randomized initialization of the two matrices W and H in the NMF decomposition. The reason is because not all the bases labelled as percussive from NMF are purely percussive so, non-percussive sounds are also synthesized as percussive sounds by the proposed method. Taking into account mixtures of percussive and harmonic instruments and singing-voice, the proposed method improves the separation performance compared with the other methods. Results show that an advantage of the proposed method, which is not exhibited by the other compared methods, is the capability to successfully remove the singing voice sounds in the separated percussive signal.

Future work will be focused on two topics. First, we

 ISBN: 978-1-61208-487-9

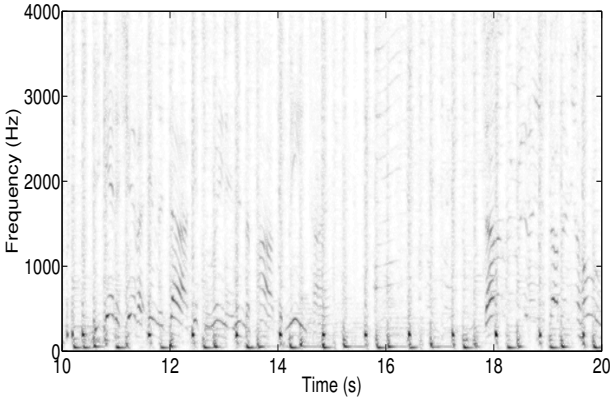

Figure 9. Percussive separation of the method MFS evaluating the interval [10s-20s] of the file M8 of the database T2. Higher grey level represents higher energy of each frequency

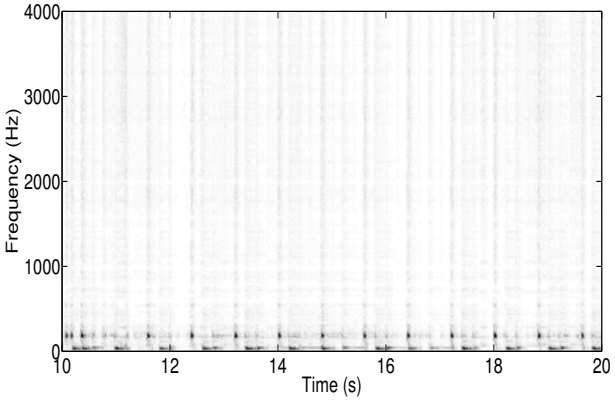

Figure 10. Percussive separation of the proposed method evaluating the interval [10s-20s] of the file M8 of the database T2. Higher grey level represents higher energy of each frequency

will investigate smart initializations based on properties of percussive sounds to improve the quality of the percussive separation. Second, new measures to discriminate the rhythmic accompaniment will be investigated (e.g., bass line).

ACKNOWLEDGMENT

This work was supported by the Andalusian Business, Science and Innovation Council under project P2010- TIC-6762 (FEDER) and the Spanish Ministry of Economy and Competitiveness under Projects TEC2012-38142-C04-01, TEC2012-38142-C04-03 and TEC2012-38142-C04-04.

REFERENCES

[1] N. Ono, K. Miyamoto, H. Kameoka, and 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), 2008, pp. 139–144.

[2] 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.

[3] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in Proceedings of Digital Audio Effects (DAFX), 2010, pp. 1–4.

[4] 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.

[5] J. Becker and C. Rohlfing, "A segmental spectral flatness measure for harmonic-percussive discrimination," in Proceedings of International Student Conference on Electrical Engineering, 2013, pp. 1–4.

[6] J. Becker, C. Sohn, and C. Rohlfing, "NMF with spectral and temporal continuity criteria for monaural sound source separation," in Proceedings of European Signal Processing Conference (EUSIPCO), 2014, pp. 316–320.

[7] 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, 2014, pp. 1–17.

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

[9] 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, 2013, pp. 2096–2107.

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

[11] "D. Ellis, Chroma features analysis and synthesis," 2007, URL: http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/ [accessed: 2016-01-02].

[12] A. Gray and J. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 22, no. 3, 1974, pp. 207–217.

[13] "Activision," 2016, URL: https://es.wikipedia.org/wiki/Guitar_Hero_5 [accessed: 2016-02-02].

[14] "Activision," 2016, URL: https://en.wikipedia.org/wiki/Guitar_Hero_World_Tour [accessed: 2016-02-02].

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

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

[17] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll, "Blind enhancement of the rhythmic and harmonic sections by nmf: Does it help?" in Proceedings of the 35th German Annual Conference on Acoustics, Acoustical Society of the Netherlands, 2009, pp. 361–364.

[18] Y. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in Proceedings of the 14th International Society for Music Information Retrieval (ISMIR) Conference, 2013, pp. 1–6.

[19] "Audio demo," 2016, URL: https://dl.dropboxusercontent.com/u/22448214/WebSIGNAL2016/indexSIGNAL2016.html [accessed: 2016-04-28].

 ISBN: 978-1-61208-487-9