Drum extraction in single channel audio signals using multi-layer non negative matrix factor deconvolution

Clément Laroche, Hélène Papadopoulos, Matthieu Kowalski, Gael Richard

DRUM EXTRACTION IN SINGLE CHANNEL AUDIO SIGNALS USING MULTI-LAYER NON NEGATIVE MATRIX FACTOR DECONVOLUTION

Clément Laroche[⋆†] *Hélène Papadopoulos*[†] *Matthieu Kowalski*[†‡] *Gaël Richard*[⋆]

[⋆] Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, Paris, France
[†] Univ Paris-Sud-CNRS-CentraleSupelec, LSS, Gif-sur-Yvette, France
[‡] Parietal project-team, INRIA, CEA-Saclay, France

ABSTRACT

In this paper, we propose a supervised multilayer factorization method designed for harmonic/percussive source separation and drum extraction. Our method decomposes the audio signals in sparse orthogonal components which capture the harmonic content, while the drum is represented by an extension of non negative matrix factorization which is able to exploit time-frequency dictionaries to take into account non stationary drum sounds. The drum dictionaries represent various real drum hits and the decomposition has more physical sense and allows for a better interpretation of the results. Experiments on real music data for a harmonic/percussive source separation task show that our method outperforms other state of the art algorithms. Finally, our method is very robust to non stationary harmonic sources that are usually poorly decomposed by existing methods.

***Index Terms*—** Drum extraction, Source separation, Non-negative matrix factorization

1. INTRODUCTION

The decomposition of audio signals in terms of elementary atoms is a very active field of research. Rank reduction methods use a dictionary of atoms and decompose a signal using a small number of them [1, 2] (usually much less than the dimension of the original space): thus they significantly reduce the amount of information that is necessary to represent the signal. These so called factorization methods rely on the fact that audio data is largely redundant as it often contains multiple correlated versions of the same physical event (note, drum hits, ...) [3]. Principal Component Analysis [4], Independent Component Analysis [5] or Non-negative Matrix Factorization (NMF) [2] produce such decompositions.

In audio signal processing, NMF has been extensively used for many different tasks such as source separation [6, 7, 8] or automatic transcription [9, 10]. However, a classic NMF does not lead to satisfactory results. It is often necessary to constrain the decomposition [11] or to rely on physical models of specific instruments [7] in order to obtain satisfying results. The goal of NMF is to approximate a data matrix $V \in \mathbb{R}_+^{f\times t}$ as $V \approx \tilde{V} = WH$, with $W \in \mathbb{R}_+^{f\times k}$ and $H \in \mathbb{R}_+^{k\times t}$ and where k is the rank of the decomposition, typically chosen such that $k(f+t) << ft$ (i.e., WH is a compressed version of V). In audio and signal processing, V is usually a spectrogram, W is a *dictionary* or a set of *patterns* and H contains the *activation coefficients*. For a stationary source, the NMF decomposition is quite accurate as each pattern of W will code an individual note. However, in the case of non stationary sources (drums, voice...), the NMF will split the audio events (drum hit, note) in multiple locally invariant patterns that do not have physical sense. This factorization of a single audio event by a multitude of templates in W makes the interpretation of the results difficult and makes it harder to pinpoint the issues of a specific method. This is particularly true for the task of Harmonic/Percussive Source Separation (HPSS) [12] in the NMF framework where the percussive part of the signal is mainly composed of non stationary sounds.

In order to represent correctly non stationary sounds in the NMF framework, Smaragdis introduced the Non-negative Matrix Factor Deconvolution (NMFD) [13]. The matrix W is a Time-Frequency (TF) atom where the spectral content may vary over time. This method was used with great success on drum only signals to perform automatic drum transcription [14] and drum playing technique identification [15]. However, these methods see their performance drop when the drum is part of a multi-instruments mix.

In our previous work on HPSS [12, 16] a NMF decomposes the audio signal in sparse orthogonal components (using the Projected NMF [17]) that are well suited for representing the harmonic part, while the percussive part is represented by a regular nonnegative matrix factorization decomposition constrained by a dictionary. The dictionary is created by performing a NMF on a large drum database. The limiting point of this method is that a drum hit is represented by a multitude of templates thus the activation of the percussive part does not have a strong physical sense as the templates do not correspond to real percussive events.

In this paper we develop a Multi-Layer NMFD suitable for HPSS that allows a better representation of the drums. More precisely, the NMFD is constrained using fixed drum dictionaries to specifically extract the drum part while the other instruments are extracted by sparse orthogonal components. The decomposition of the percussive part has more physical sense as the template activation corresponds to a real percussive even which allows for an easier interpretation of the results. The merit of this new method is experimentally demonstrated for an HPSS task on a database of 89 real music songs. Our method revealed to be particularly interesting as it is able to extract the drum more effectively than other state of the art methods. Another strong point of our method is that it is able to extract the voice mainly in the harmonic part whereas other state of the art method split the voice in the harmonic and percussive part.

The paper is organized as follows. In Section 2, the NMFD is described and we present a small overview of its different applications. The Multi-Layer NMFD is then introduced with its setup process in Section 3. We detail our experimental protocol and the results obtained on real audio signals in Section 4. Finally, some conclusions

H. Papadopoulos is supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Program

This work was supported by a grant from DIGITEO

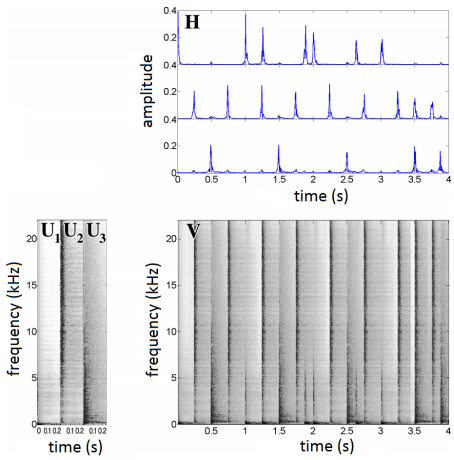

Fig. 1: NMFD decomposition of the spectrogram V. The top matrix is H and the left matrix is $W = [U_1 U_2 U_3]$

are drawn in Section 5.

2. NON NEGATIVE MATRIX FACTOR DECONVOLUTION

The NMFD is proposed as an extension of the NMF in [13]. The NMFD factorizes V in TF patterns as:

$$V \approx \tilde{V} = \sum_{\tau=0}^{L-1} W_\tau \overleftarrow{H}^{\tau} \tag{1}$$

where L is the length of a time/frequency pattern, W_τ is a learned set of arrays indexed by τ (which may be seen as a tensor), H are the activation coefficients of the patterns and $\overleftarrow{*}^{\tau}$ is a column shift operator defined in [13]: Equation (1) is thus a convolution operation. It may be interesting for interpretation purposes to rewrite the expression of the tensor of atoms as follows: $[U_k]_{f\tau} = [W_\tau]_{fk}$. The U_k matrix is a two-dimensional atom and k corresponds to a musical time/frequency event where the spectral content may vary over time.

Figure 1 shows the decomposition of a spectrogram V (bottom right) containing percussive elements (drum loop). In this example, U_1, U_2 and U_3 (bottom left) are the three learned TF atoms, each corresponding to a drum element (respectively, bass, hi-hat and snare). In order for this decomposition to make sense, the activations must be sparse and the TF pattern represent an entire musical event. A sparsity constraint term is generally added to promote this property.

A limitation of this method is that it does not model the variations between different occurrences of the same event: its duration and the evolution of its spectral content are fixed.

This type of methods was also directly used in the time domain using an algorithm of semi- NMF [18]. It estimates the waveforms of the various components of the signal.

3. MULTI-LAYER NMFD

In this section we present a model of a multi-layer decomposition where the NMFD decomposes the percussive non-stationary part of the signal while the harmonic instruments are extracted by a sparse orthogonal part.

3.1. Multi-Layer decomposition

Our method for HPSS differs from previous methods as we are going to build a multi-layer decomposition where the percussive part is extracted using NMFD with a fixed W_τ. The templates used are created using real drum sounds. Previous methods using NMFD focused on drum transcription.

In our previous work on HPSS, the harmonic components are represented by the Projective NMF (PNMF) while the percussive ones are represented by a regular NMF decomposition as follow:

$$V \approx \tilde{V} = W_P H_P + W_H W_H^T. \tag{2}$$

In the present work, in order to improve the drum extraction in the case of a multi-instruments mixture, the percussive components are modeled using NMFD. Most of the harmonic components are represented by the sparse orthogonal term $W_H W_H^T$ while the percussive ones are extracted in the NMFD components. Let V be the magnitude of the Short Time Fourier Transform (STFT) of the input data. The model called Multi-Layer NMFD (ML-NMFD) is then given by

$$V \approx \tilde{V} = \sum_{\tau=0}^{L-1} W_{\tau,P} \overleftarrow{H_P}^{\tau} + W_H W_H^T V, \tag{3}$$

Here $W_{\tau,P}$ is fixed so our objective is to find a set H_P to approximate the percussive part while the harmonic part is extracted by the orthogonal components. The model is optimized by reducing the value of the cost function between the estimated matrix $\tilde{V}$ and the data matrix V. Let consider the optimization problem

$$\min_{H_P, W_H} D_{KL}(V|\tilde{V}) = \frac{V}{\tilde{V}} - \log \frac{V}{\tilde{V}} - 1, \tag{4}$$

where D_{KL} is the Kullback-Leibler (KL) divergence and $\frac{*}{*}$ is the element wise division. We decided to use the KL divergence as it yields better results in our experiments. The Itakura-Saito (IS) divergence is scale invariant, it means that the low energy components of the spectrogram bear the same relative importance as the higher ones. The small discrepancies between the learned dictionary and the target drum impact the results more in the case of the IS divergence. Using the same optimization scheme as [19], we obtain the following multiplicative update rules and at each step the objective function is guaranteed non-increasing:

$$W_H \leftarrow W_H \otimes \frac{(ZV^T W_H) + (VZ^T W_H)}{(\phi V^T W_H) + (V\phi^T W_H)} \tag{5}$$

with

$$Z = V \otimes \tilde{V}^{-1} \qquad \phi = I \otimes \tilde{V}^{-1},$$

and $I \in \mathbb{R}^{f \times t}; \forall i,j \quad I_{i,j} = 1$, where $(*)^{-1}$ and $\otimes$ are respectively element wise matrix power and multiplication.
Finally for H_P, we obtain:

$$H_P \leftarrow H_P \otimes \frac{\sum_\tau W_{\tau,P}^T (V \overleftarrow{\tilde{V}^{-1}}^{\tau})}{\sum_\tau W_{\tau,P}^T \overleftarrow{I}^{\tau}}. \tag{6}$$

As $W_{\tau,P}$ is fixed, we then update the two variables W_H and H_P sequentially until the convergence is reached. Not having to update the matrix $W_{\tau,P}$ speeds up the optimization.

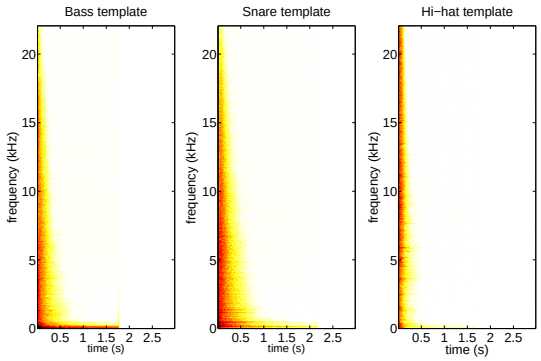

Fig. 2: Sample of the dictionary for the percussive part extracted from the ENST-Drum database [20].

3.2. Construction of the dictionaries

$W_{\tau,P}$ is a fixed drum dictionary built using isolated drum hits from the ENST-Drums database [20]. Each of the W atoms corresponds to a particular drum element. Figure 2 shows three TF atoms, each corresponding to an element of drum. In order to keep the computation time reasonable, we decided to consider only single-drum strikes from 6 different elements of the drum (bass drum, snare, hihat open, hithat closed, ride cymbal, crash cymbal).

3.3. Signal reconstruction

3.3.1. Via Wiener filtering

The percussive signal $x_p(t)$ is synthesized using the magnitude percussive spectrogram $V_P = \sum_{\tau=0}^{L-1} W_{\tau,P} \overleftarrow{H}_P^{\tau}$. To reconstruct the phase of the STFT of the percussive part, we use a Wiener filter [21] to compute a percussive mask as:

$$\mathcal{M}_P = \frac{V_P^2}{V_H^2 + V_P^2}. \tag{7}$$

To retrieve the percussive signal as:

$$x_P(t) = \text{STFT}^{-1}(\mathcal{M}_P \otimes X). \tag{8}$$

Where X is the complex spectrogram of the mixture and STFT^{-1} is the inverse transform of the STFT. We retrieve the harmonic signal using the same technique.

3.3.2. Drum synthesis

Because the dictionary are built using real audio sounds of drum strikes, we have access to the complex spectrogram of the dictionary as $|\mathcal{X}_{\tau,P}| = W_{\tau,P}$. We can reconstruct a synthetic drum signal using the equation:

$$x_P(t) = \text{STFT}^{-1}[\sum_{\tau=0}^{L-1} \mathcal{X}_{\tau,P} \overleftarrow{H}_P^{\tau}]. \tag{9}$$

This method produces a sound with no interference (i.e., the sound is synthesized using only real drum sounds) that could be used to restore a drum audio file that is damaged or not recorded in the optimal conditions. There is no objective means to evaluate the synthesis quality, however the sound examples provided in the companion page [1] show the potential of this alternative reconstruction scheme.

[1] http://perso.telecom-paristech.fr/ laroche/Article/ICASSP2017/index.html

4. STATE OF THE ART COMPARISON

In this section we compare the proposed ML-NMFD with other state of the art methods on an HPSS task.

4.1. Experimental protocol

Using the experimental protocol from [22, 11], we perform a blind source separation evaluation on the 89 audio signals from the database Medley-dB [23] that contain percussive/drum sounds. We run two tests on the whole database. In the first test, the vocal part is omitted as in [11]. In the second test, we include the voice in the separation test. We compare the ML-NMFD to three other recent state of the art methods: Constrained NMF (CoNMF) [11], Median Filtering (MF) [24] and Structured Projective NMF (SPNMF) [22]. MF and CoNMF are two unsupervised methods while SPNMF is a supervised method that uses drum dictionaries learned by NMF. CoNMF is our own re-implemented algorithm and the MF implementation is taken from [25]. Compared to other state of the art methods, the proposed ML-NMFD algorithm is the most computationally intensive. For a 30 s signal our method takes 2 min to converge compared to approximately 30 s for SPNMF and CoNMF while MF is almost instantaneous. All the signals are sampled at 44.1 kHz. We compute the STFT with a 2048 sample-long Hann window and a 50% overlap. The harmonic and percussive signals are reconstructed using Wiener filtering (see 3.3.1). The results are compared using the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Artifact ratio (SAR) of the harmonic and percussive sources using the BSS Eval toolbox [26]. Each box-plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles while the whiskers indicate the minimum and maximum values.

4.2. Results without the vocal part

The results of the separation on the 89 songs of the Medley-dB database are displayed in Figure 3. The SDR results for the harmonic, percussive and overall separation are higher with the proposed ML-NMFD algorithm even if the database [23] contains a wide variety of percussive instruments (tabla, tambourine, timpani and other) that are not in the training database. The SIR scores show the strength of our method (minimum +4dB for the overall reconstruction). Having TF dictionaries that keep the temporal information of the drum sounds considerably improves the separation results compared to the results of SPNMF. Our method obtains similar results for the SAR as the MF and a lower SAR than SPNMF and CoNMF. Because we are using the STFT of a drum strike as a template, it produces more artefacts in the extracted source as the dictionary and the target drum are different.

4.3. Results with the vocal part

In this section we address the problem of drum extraction when a singing voice is present in the mix. Singing voice contain both quasi-harmonic components (such as vowels) and explosive and non-stationary components such as plosive (“p”, “k”, ...). Also singing voice produces some very non-stationary sounds (i.e., vibratos, tremolos, bends), thus the voice cannot be labeled as a percussive nor harmonic instrument. Most state of the art methods split the voice in the harmonic and percussive parts. That is why many methods that rely on HPSS and more precisely MF [24] as a

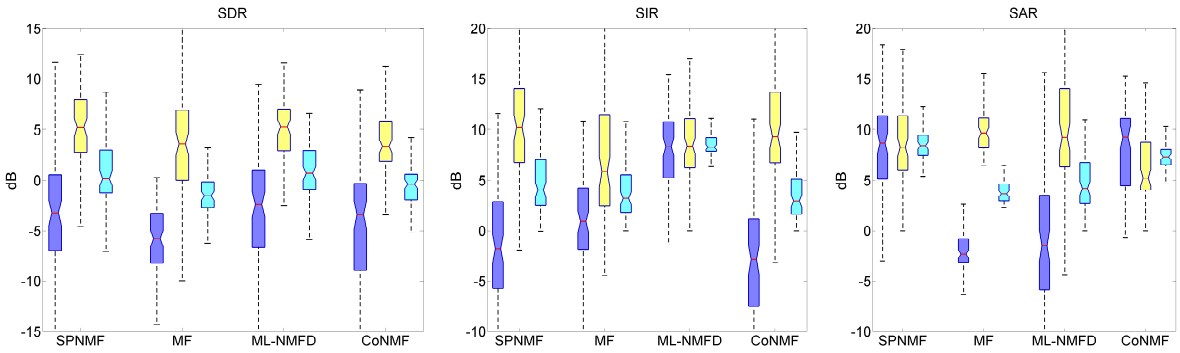

Fig. 3: SDR, SIR and SAR for percussive (left/blue), harmonic (middle/yellow), mean (right/cyan) separation results on the database for the four methods.

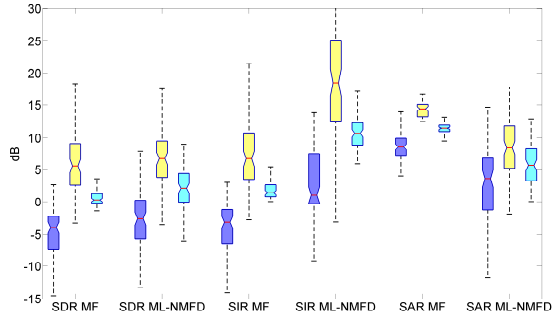

Fig. 4: SDR, SIR, SAR results for percussive (left/blue), harmonic (middle/yellow), mean (right/cyan) separation results on the database for the two methods.

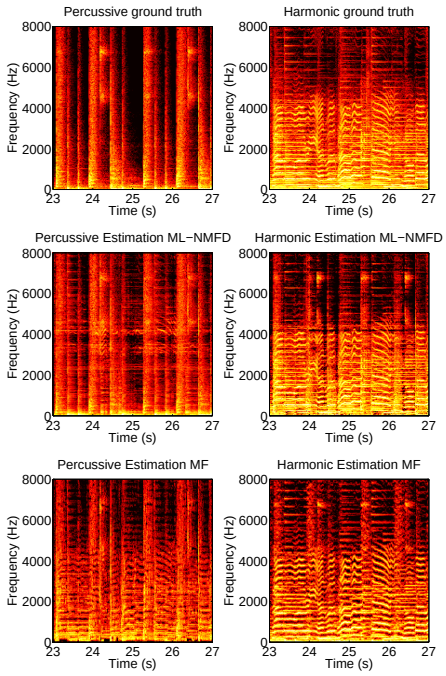

Fig. 5: Comparaison results between the ground truth (top) ML-NMFD (middle) and HPSS (bottom).

preprocessing step still need to refine the output of the decomposition. Melody extraction [27] algorithms are much more efficient if the influence of the percussive sources is diminished in the mix. Similarly, drum transcription algorithms [28] are more accurate if the harmonic instruments are not part of the signal and could obtain some improvement if the transient part of the voice was completely removed. For this reason we decided in our test to classify the voice as an harmonic instrument. Figure 4 shows the results of the HPSS task on the 89 songs of the Medley-dB database. The ML-NMFD outperforms the MF. The results shows that the separation scores of the ML-NMFD are higher for SDR and SIR. The decomposition creates less interferences and the voice is extracted by the harmonic part (i.e., SIR +5dB for the two layers). For the MF algorithm, the strongly non stationary components of the voice are extracted in the percussive part while the more linear components are extracted in the harmonic part which results in a low SIR for the overall decomposition. In Figure 5 we can see that the MF has much more harmonic residual than the ML-NMFD in the percussive decomposition which confirm the conclusion from previous results.

5. CONCLUSION

In this article, we have shown that the proposed ML-NMFD is a robust and effective method to extract the drums in a multi-instruments single channel audio file. The ML-NMFD algorithm outperforms the other state of the art methods by a significant margin. Finally, in the case of audio signals with vocal parts, the ML-NMFD is able to extract the drums much more effectively than the classical MF methods while the vocal track is extracted in the harmonic part. This method could be a powerful pre-processing step for other algorithms as explained in Section 4.3.

Future work will focus on using this method for drum transcription, drum re-synthesis (see Section 3.3.2) and drum remixing. Sound examples of the different tests can be found on the companion page.

REFERENCES

[1] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[3] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Transactions on Signal Processing*, vol. 28, no. 2, pp. 27–38, 2011.

[4] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, vol. 46, John Wiley & Sons, 2004.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, Language Processing.*, pp. 1066–1074, 2007.

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

[8] J. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, Language Process.*, pp. 564–575, 2010.

[9] S. Ewert, B. Pardo, M. Müller, and M. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

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

[11] F. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1–17, 2014.

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

[13] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. of IEEE ICA*, pp. 494–499. 2004.

[14] A. Roebel, J. Pons, M. Liuni, and M. Lagrange, "On automatic drum transcription using non-negative matrix deconvolution and Itakura Saito divergence," in *Proc. of IEEE ICASSP*, 2015, pp. 414–418.

[15] C. Wu and A. Lerch, "On drum playing technique detection in polyphonic mixtures," in *Proc. of ISMIR*, 2016.

[16] C. Laroche, H. Papadopoulos, M. Kowalski, and G. Richard, "Genre specific dictionaries for harmonic/percussive source separation," in *Proc. of ISMIR*, 2016.

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

[18] J. LeRoux, A. Cheveigné, and L. Parra, "Adaptive template matching with shift-invariant semi-NMF," in *Proc. of NIPS*, 2009, pp. 921–928.

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

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

[21] A. Liutkus and R. Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Proc. of IEEE ICASSP*, 2015, pp. 266–270.

[22] C. Laroche, M. Kowalski, H. Papadopoulous, and G. Richard, "A structured nonnegative matrix factorization for harmonic/percussive source separation," *Submitted to IEEE Transactions on Acoustics, Speech and Signal Processing*.

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

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

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

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

[27] F. Rigaud and M. Radenen, "Singing voice melody transcription using deep neural networks," in *Proc. of ISMIR*, 2016.

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