GENRE SPECIFIC DICTIONARIES FOR HARMONIC/PERCUSSIVE SOURCE SEPARATION

Clément Laroche Affiliation1 author1@ismir.edu
Second author Affiliation2 author2@ismir.edu
Third author Affiliation3 author3@ismir.edu
Fourth author Affiliation4 author4@ismir.edu

ABSTRACT

The abstract should be placed at the top left column and should contain about 150-200 words.

1. INTRODUCTION

Source separation is field of research that seeks to separate the components of an audio signal present in a record. Such separation has many applications in music : noise suppression [4] (if a source is a noise) , up-mixing [8](spatialization of the sources) or automatic transcription [2] (it is easier to operate on single source). The task is difficult due to the complexity and the variability of the music mixtures. Most datasets used for Blind Source Separation (BSS) research are small in size and they do not allow for a thorough comparison of the algorithms. Using a larger database is crucial to benchmark the separation algorithms in order to obtain a true evaluation rather than particular case results.

The large variety of audio signal can be classified into different musical genres [19]. Genres are labels created and used by humans for categorizing and describing music. They have no strict definitions and boundaries but particular genre share certain characteristics typically related to the instrumentation, rhythmic structure, and pitch content of the music and the resemblance between two pieces of music have been used to perform chord transcription [14, 17] or for downbeat detection [9]. Finally when the genre information is not available, it is possible to perform automatic genre classification [15].

In the context of BSS, Non-negative Matrix Factorization (NMF) is a widely used method for source separation. The goal of NMF is to approximate a data matrix $V \in \mathbb{R}_+^{n\times m}$ as $V \approx \tilde{V} = WH$ with $W \in \mathbb{R}_+^{n\times k}$, $H \in \mathbb{R}_+^{k\times m}$ and where k is the rank of factorization [12]. In audio signal processing, the input data is usually a Time-Frequency (TF) representation such as a short time Fourier transform (STFT) or a constant-Q transform spectrogram. Blind source separation is a difficult problem and the plain NMF decomposition does not provide satisfying results. To perform a satisfying decomposition, it is necessary to exploit various features that make each sources distinguishable from one another. Supervised algorithms in the NMF framework exploit training data or prior information in order to guide the decomposition process. For example information from the scores or from midi signals [7] can be used to initialize the learning process. The downside of this approach is that it requires well organized prior information that is not always available. Another supervised method consists in performing prior training on specific databases. For example a dictionary matrix W_{train} is learned from a big database in order to separate an instrument [10, 20]. A common method to build a dictionary for NMF is to perform a decomposition on a large training set. After the convergence, the W matrix from the decomposition is used as the dictionary matrix W_{train} in the separation [10]. Another method is detailed in [20], a dictionary matrix is created by extracting template spectra from isolated drum samples. The dictionary is then used in a NMF decomposition to perform drum transcription. This method requires minimum tuning from the user. However, the dictionary should match the target instrument for satisfying performances.

In this paper, we focus on the task of harmonic/percussive source separation (HPSS) using the method developed in [11]. We adapt the method to be used with a drum dictionary to extract the percussive instruments. This method is explained in detail in the preprint. The problem of using a fixed dictionary matrices is that within a database, the same instrument can sound differently depending on the recording condition and post processing treatment. In order to represent correctly one instrument, ones can decide to learn a dictionary on a large database. However, the problem of over-fitting the data exist. In order to overcome this problem and to build effective dictionaries we decided to use genre specific training data. As they share similar features, genre specific information can provide an insight on the structure of the audio signal. The main contribution of this article is that we developed a genre specific method to build a drum NMF dictionary that obtains consistent results on a HPSS task. By using genre specific dictionary we were able to improve the separation score compared to a universal dictionary.

2. STRUCTURED PROJECTIVE NMF (SPNMF)

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

2.1 Presentation of the orthogonal and projective NMF

The aim of PNMF is to find a non negative projection matrix $P \in \mathbb{R}_+^{n\times n}$ such that $V \approx \tilde{V} = PV$. In [22] Yuan & al. propose to seek P as an approximative projection matrix under the form $P = WW^T$ with $W \in \mathbb{R}_+^{n\times k}$ with $k \leqslant n$. The PNMF problem reads :

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

PNMF is similar to the NMF problem and can be simply obtained by replacing the activation matrix H by W^TV. It is shown in [21] that the PNMF gives a much sparser decomposition than NMF.

Another very similar approach is the ONMF [6]. It consists in solving the following problem:

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

In this method, orthogonality between nonnegative basis functions is enforced during the optimization process. In theory, it seems that PNMF and ONMF lead to similar decompositions, as the W matrix estimated by PNMF is almost orthogonal (i.e., $\|W^TW - I_k\|^2$ is small). However in practice, enforcing the orthogonality between the base at every iteration is a constraint too strong to decompose audio signal [11].

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

2.2 Principle of the SPNMF

The orthogonal basis functions of PNMF are not flexible enough to decompose a complex audio signal. As stated in [5], harmonic instruments have sparse basis functions whereas percussive instruments have much flatter spectra. As the columns of W are orthogonal, when two sources overlap in the Time-Frequency (TF) plane only one basis function will represent the mixture which is not adequate for efficient separation. To overcome this problem, we propose to add a standard NMF decomposition term to the PNMF. We can expect that most of the harmonic components will be represented by the orthogonal part while the percussive ones will be the regular NMF components. Using a similar model as in our preliminary work [11], let V be the magnitude spectrogram of the input data. The model is then given by

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

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

$$V \approx \tilde{V} = W_HW_H^TV + W_PH_P. \tag{4}$$

The data matrix is approximated by an almost orthogonal sparse part that codes the harmonic instruments $V_H = W_HW_H^TV$ and a non constrained NMF part that codes the percussive instruments $V_P = W_PH_P$. We use here a fixed drum dictionary W_p in the percussive part of the SPNMF as a fully unsupervised SPNMF model does not allow for a satisfying harmonic/percussive source separation [11].

2.3 Algorithm Optimization

In order to obtain such a decomposition, we can use a measure of fit $D(x|y)$ between the data matrix V and the estimated matrix $\tilde{V}$. $D(x|y)$ is a scalar cost function and in this article, we use the Itakura Saito (IS) divergence.

The SPNMF model gives the cost function :

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

A solution of this problem can be obtained by iterative multiplicative update rules following the same strategy as in [13, 22] which consists in splitting the gradient with respect to (wrt) one variable (here W_H for exemple) $\nabla_{W_H}D(V|\tilde{V})$ in its positive $[\nabla_{W_H}D(V|\tilde{V})]^+$ and negative parts $[\nabla_{W_H}D(V|\tilde{V})]^-$. The multiplicative updates for SPNMF are then given by:

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

where $\otimes$ is the Hadamard product or element-wise product. The SPNMF algorithm with a fixed dictionary matrix is:

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

Algorithm 1: SPNMF with the drum dictionary matrix.

2.4 Signal reconstruction

The percussive signal $x_p(t)$ is synthesized using the magnitude percussive spectrogram $X_P = W_PH_P$. To reconstruct the phase of the percussive part, we use a generalized Wiener filter [16] to create a percussive mask as:

$$\mathcal{M}_P = \frac{X_P^2}{X_M^2 + X_P^2}. \tag{6}$$

To retrieve the percussive signal as,

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

Where X is the complex spectrogram of the mixture. Similarly for the harmonic part, we obtain:

$$\mathcal{M}_H = \frac{X_H^2}{X_M^2 + X_P^2}, \quad (8)$$

and:

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

3. CONSTRUCTION OF THE DICTIONARY

3.1 Optimal size for the dictionary

The first step to build a NMF drum dictionary is to select the rank of factorization. We run different tests on the public SiSec database from [1]. It is composed of polyphonic real-world music excerpts and each music signal contains percussive, harmonic instruments and vocals. The duration of the four recording is ranging from 14 to 24 s. The goal is to perform an harmonic/percussive decomposition. Following [5], we will not consider the vocal part and we will build mixture signals only from the percussive and harmonic instruments. All the signals are sampled at $44.1kHz$. We compute the STFT with a and 2048 sample-long Hann window with a 50% overlap.

The drum signal used for the training comes from the database [18] and the signal is around 3min long. We used 14 files from the database were the drummer is playing a *drum phrase*. We compute an NMF decomposition with different rank of factorization ($k = 12$, $k = 50$, $k = 100$, $k = 500$, $k = 1000$ and $k = 2000$) on the drum signal alone to obtain 6 drum dictionaries. The dictionaries are then used to perform a HPSS on the four songs of the SiSEC database using the SPNMF algorithm. The results are displayed on figure 1.

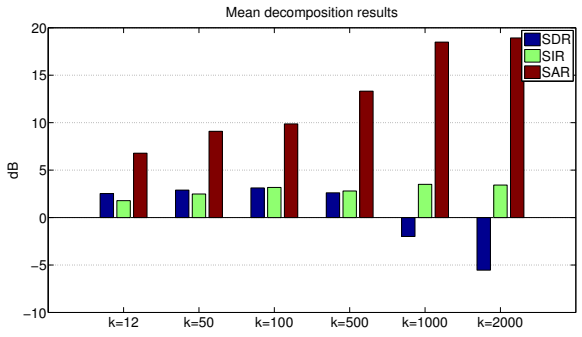

Figure 1: Average results on the SiSec database.

The optimal value is reached for $k = 100$, then the SDR decrease rapidly for $k \geqslant 500$. The high value of SAR (for $k \geqslant 500$) are explained because the separation process is not satisfying. The harmonic signal given at the end of the algorithm is composed of most of the original signal therefore it cause the SAR to be very high.

We can conclude that for a 3min drum signal, the optimal dictionary size for the SPNMF algorithm is 100.

Classical	JoelHelander Definition MatthewEntwistle AnEveningWithOliver MusicDelta Beethoven
Electronic/Fusion	EthanHein 1930sSynthAndUprightBass TablaBreakbeatScience Animoog TablaBreakbeatScience Scorpio
Jazz	CroqueMadame Oil MusicDelta BebopJazz MusicDelta ModalJazz
Pop	DreamersOfTheGhetto HeavyLove NightPanther Fire StrandOfOaks Spacestation
Rock	BigTroubles Phantom Meaxic TakeAStep PurlingHiss Lolita
Singer/Songwriter	AimeeNorwich Child ClaraBerryAndWooldog Boys InvisibleFamiliars DisturbingWildlife
World/Folk	AimeeNorwich Flying KarimDouaidy Hopscotch MusicDelta ChineseYaoZu
Non specific	JoelHelander Definition TablaBreakbeatScience Animoog MusicDelta BebopJazz DreamersOfTheGhetto HeavyLove BigTroubles Phantom AimeeNorwich Flying MusicDelta ChineseYaoZu

Table 1: Song selected for the training database.

3.2 Database

The dataset is taken from medley-dB [3], it is composed of polyphonic real-world music excerpts. It has 122 music signals and 87 of them contain percussive instruments, harmonic instruments and vocals. The signals that do not contain a percussive part are not part of the evaluation. We will be using the song of the genre, *Classical* (8 songs), *Singer/Songwriter* (17 songs), *Pop* (10 songs), *Rock* (20 songs), *Jazz* (11 songs), *Electronic/Fusion* (13 songs) and *World/Folk* (6 songs). Because the notion of genre is quite subjective, the medley-dB database uses general genre labels. These labels should not be considered to be "precise" genre labels. There are many instances where a song could have fallen in multiple genres, and the choices were made so that each genre would be as acoustically homogeneous as possible. As we are only working with the instrumental part of the song with the "Pop" label (for example) are similar to the "Singer/Songwriter". We use 3 song of each genre to use as a training database. The songs used for the training part are not part of the evaluation. To compare the genre specific dictionary, we build a non specific/universal dictionary built using one half of one song of each genre. The files selected are:

Genre	Length(min)
Classical	22.06
Electronic/Fusion	18.66
Jazz	10.96
Pop	12.53
Rock	11.43
Singer/Songwriter	9.36
World/Folk	9.53
Non Specific (Mix)	11.03

Table 2: Length of the genre specific database.

3.3 Genre specific dictionaries

The NMF model is:

$$V \approx \tilde{V} = WH. \tag{10}$$

If V is the power spectrum of a drum signal, The matrix W is a *dictionary* or a set of *patterns* that codes the frequency information of the drum. Building a dictionary specific to an instrument that performs well on a large database is a complicated problem. Here we build genre specific drum dictionary using the medley-dB database. Using dictionary specific to the genre of music allows us to have smaller dictionaries that a more specific to the signal to decompose. It grants us lower computation time and better separation score The dictionary are build has follow. For every genre specific database of the training database, we perform and NMF with $k = 300$ on the drum signals (with the results from Section 3.1, we choose $k = 100$ per song for the NMF). Then the dictionaries are used in the SPNMF algorithm as the matrix W_P (see algorithm 1).

4. RESULTS

In this section we present the results of the algorithm with the genre specific dictionaries on the 69 song from the Medley-dB database. We perform an HPSS on the audio files using the SPNMF with the 8 dictionaries created on Section 3.3. The results on the songs are then sorted by genre and the average results are display using box-plot. 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.

The Figure **??**, 4 and 3 shows the SDR, SAR and SIR results for all the dictionaries on the *Pop* subsection. It gives us a overall idea on how all the dictionaries perform on the same database. The results using the *Pop* dictionary has the highest SDR results. The non specific dictionary is not performing as well the pop dictionary, it allows us to say that on this *Pop* database, the genre specific dictionary is giving relevant information to the algorithm. As stated in Section 3.2, some genre are similar to other and that can explain why the *Rock* and *Singer* dictionaries are giving good results too. And interesting result to note is that compared to the non specific dictionary, the genre specific *Pop* dictionary has a lower variance. The genre information allows for a higher robustness of the decomposition results.

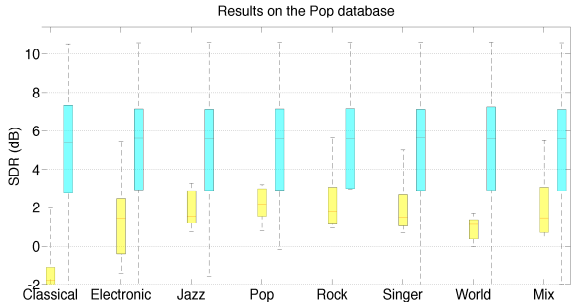

Figure 2: Results on the Pop sub-database using the SPNMF with the 8 dictionaries.

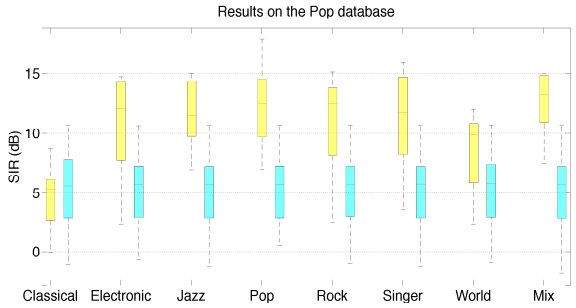

Figure 3: Results on the Pop sub-database using the SPNMF with the 8 dictionaries.

On Table 4, we display the mean separation score for all the genre specific dictionaries compared to the non specific dictionary. The genre specific dictionaries outperform the universal dictionary by a considerable margin.

4.1 Discussion

On the database *Singer/Songwriter*, *Pop*, *Rock*, *Jazz* and *World/Folk*, learning a genre specific dictionary outperform the universal dictionary. The similar pitch content of the music of the same genre is not altered by the NMF decomposition and this information improve the separation. The database *Classical* and *Electronic/Fusion* are composed of songs where the drum is only playing for a few moments. Similarly on some songs of the *Electronic/Fusion* database, the electronic drum reproduces the same pattern during the whole song making the drum part very redundant. Because of these two factors, the genre specific dictionaries are not performing correctly.

The main drawback of using a NMF dictionary is that the decomposition is not unique and can any permutation is also a solution. Because of that, we lose the temporal structure of the original drum signal.

5. CONCLUSION

Using genre specific information in order to build more relevant drum dictionaries was proven to be a powerful method to improve the HPSS. The dictionaries still have an imprint of the genre even after the NMF decomposition.

Genre	Classical	Electronic/Fusion	Jazz	Pop	Rock	Singer/Songwriter	World/Folk
Genre specific Results(dB)							
SDR	2.95	**0.54**	5.94	**3.76**	**0.97**	**4.69**	**1.83**
SIR	**9.41**	8.51	**11.1**	**9.77**	**12.0**	9.80	**8.33**
SAR	12.0	11.9	**15.6**	13.9	**19.1**	**17.2**	**18.6**
Non specific (dB)							
SDR	**3.00**	0.50	**6.02**	3.40	-0.11	3.71	0.66
SIR	9.25	**9.18**	10.5	8.72	10.6	**10.3**	7.53
SAR	**17.6**	**14.1**	15.4	**14.5**	15.4	16.9	16.4

Table 3: Mean SDR, SIR and SAR results of the Medley-dB database.

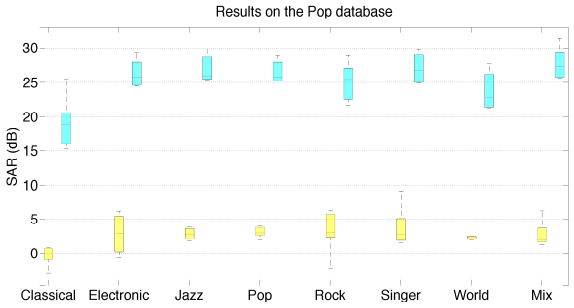

Figure 4: Results on the Pop sub-database using the SPNMF with the 8 dictionaries.

This is a first step in order to produce dictionaries capable of extracting a wide variety of audio signal.

Future work will be dedicated into taking into account the temporal structure of the original drum signal in order to maintain the temporal coherence of dictionary.

6. ANNEXE

6.1 Itakura Saito divergence

The Itakura Saito divergence gives us the problem,

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

The gradient wrt W_1 gives

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

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

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

with

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

and $I = ones(size(V))$.

Similarly, the gradient wrt W_2 gives

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

and

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

Finally, the gradient wrt H_2 gives

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

and

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

7. REFERENCES

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

[2] 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*, pages 65–68, 2007.

[3] 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.

[4] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, 1979.

[5] 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*, 2014(1):1–17, 2014.

[6] S. Choi. Algorithms for orthogonal nonnegative matrix factorization. In *Proc. of IEEE IJCNN*, pages 1828–1832, 2008.

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

[8] D. Fitzgerald. Upmixing from mono-a source separation approach. In *IEEE proc. of DSP*, pages 1–7, 2011.

[9] J. Hockman, M. Davies, and I. Fujinaga. One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass. In *In proc. of ISMIR*, pages 169–174, 2012.

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

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

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

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

[14] K. Lee and M. Slaney. Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):291–301, 2008.

[15] T. Li, M.Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *In Proc. of ACM*, pages 282–289, 2003.

[16] A. Liutkus and R. Badeau. Generalized wiener filtering with fractional power spectrograms. In *IEEE Proc. of ICASSP*, pages 266–270, 2015.

[17] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie. Using hyper-genre training to explore genre information for automatic chord estimation. In *Proc. of ISMIR*, pages 109–114, 2012.

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

[19] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[20] Chih-Wei Wu and Alexander Lerch. Drum transcription using partially fixed non-negative matrix factorization. In *Proc. of EUSIPCO*, 2008.

[21] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Network.*, 21(5):734–749, 2010.

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