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

Supervised algorithms for audio source separation often use a priori information such as musical scores and learned dictionaries to achieve a decomposition. Audio signals are diverse because of the inconsistency of the musical performance, recording condition and post processing treatment, as a result it is generally impossible to build prior information that is relevant on any audio signal. Current state-of-the-art methods for source separation use large instruments specific data to build dictionaries to take into account these variability. However, the dictionaries are still too generic and they do not characterize completely a target instrument as the large amount of information is not properly compressed. In this article, we prove that musical genre information is suitable to further specialize the dictionaries in order to improve the decomposition results. Our test on a task of harmonic/percussive source separation on a large database proves that the genre is a relevant feature as the drum dictionaries built using genre specific information perform better that an universal drum dictionary.

1. INTRODUCTION

Source separation is a field of research that seeks to separate the components of a recorded audio signal. Such a separation has many applications in music such as upmixing [11] (spatialization of the sources) or automatic transcription [40] (it is easier to work on single sources). The separation task is difficult due to the complexity and the variability of the music mixtures. Most datasets used for Blind Audio Source Separation (BASS) research are small in size and they do not allow for a thorough comparison of the source separation algorithms. Using a larger database is crucial to benchmark the different algorithms in order to get an insight of their performances rather than to focus on particular case results.

The large variety of audio signals can be classified into various musical genres [39]. 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 instrumentation, rhythmic structure, and pitch content of the music. This resemblance between two pieces of music has been used as an information to improve chord transcription [25, 29] or downbeat detection [16] algorithms. Genre information can be obtained using annotated labels. When the genre information is not available, it can be retrieved using automatic genre classification algorithms [28, 39].

In the context of BASS, Non-negative Matrix Factorization (NMF) is a widely used method. The goal of NMF is to approximate a data matrix $V \in \mathbb{R}_+^{n\times m}$ as

$$V \approx \tilde{V} = WH \tag{1}$$

with $W \in \mathbb{R}_+^{n\times k}$, $H \in \mathbb{R}_+^{k\times m}$ and where k is the rank of factorization [23]. 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 obtain a satisfying decomposition, it is necessary to exploit various features that make each source 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 can be used to initialize the learning process [9]. The downside of these approaches is that they require well organized prior information that is not always available. Another supervised method consists in performing prior training on specific databases. A dictionary matrix W_{train} can be learned from a big database in order to separate instruments [19,42]. Such methods require minimum tuning from the user. However, the dictionaries must match the target instruments to obtain satisfying performances.

Harmonic/percussive source separation has numerous applications as a preprocessing step for other audio tasks. For example most multi-pitch estimation models [21], instruments recognition [8] and melody extraction [35] algorithms are much more efficient when processing harmonic data only. Similarly, beat tracking [7] and drum transcription algorithms [32] are more accurate if the harmonic instruments are not part of the analyzed signal. The HPSS algorithm [10] can be also used as a preprocessing step to increase the performance for singing pitch extraction and voice separation [17].

In this paper, we focus on the task of harmonic/percussive source separation (HPSS) leveraging on the method developed in [12]. In this work, an unconstrained NMF decomposes the audio signal in sparse orthogonal components that are well suited for representing the harmonic component, while the percussive part is represented by a regular nonnegative matrix factorization decomposition. In [13], we have adapted the algorithm using a trained drum dictionary to improve the extraction of the percussive instruments. The problem of using a fixed dictionary matrix is that within different music pieces of a database, the same instrument can sound differently depending on the recording conditions and post processing treatments. In order to accurately represent an instrument, one can decide to learn a dictionary on a very large database but takes the risk of over-fitting the data. Furthermore when databases cover a wide variety of genres, instrumentation may strongly differ from one piece to another. In order to avoid this problem and to build effective dictionaries, we propose to use genre specific training data.

The main contribution of this article is that we develop a genre specific method to build NMF drum dictionaries that give consistent and robust results on a HPSS task. The *genre specific dictionaries* are able to improve the separation score compared to a *universal dictionary* trained from all available data.

Section 2 defines the context of our work, Section 3 presents the proposed algorithm while Section 4 describes the construction of specific dictionaries. Finally Section 5 details the results of the HPSS on 65 audio files and we suggest some conclusions in Section 6.

2. TOWARD GENRE SPECIFIC INFORMATION

2.1 Genre information

Musical genre is one of the most prominent high level music descriptors. Electronic Music Distribution has become more and more popular in recent years and music catalogues never stop to increase (the biggest online services now propose around 1 million tracks). In that context, associating a genre to a musical piece is crucial to help users finding what they are looking for. As mentioned in the introduction, genre information has been used as a cue to improve some content-based music information retrieval algorithms. If an explicit definition of musical genres is not really available [4], musical genre classification can be performed automatically [26].

Source separation has been used extensively in order to help the genre classification process [22, 34]. However, genre information has not been yet exploited to guide the decomposition process.

2.2 Methods for dictionary learning

Audio data is largely redundant in two main aspects: it often contains multiple correlated versions of the same physical event (note, drum hits...) [38]. The relevant information is generally of much reduced dimensionality compared to the original data sets hence the idea to exploit this redundancy to reduce the amount of information necessary for the representation of a musical signal.

Many rank reduction methods, such as Single Value Decomposition (K-SVD) [2], Vector Quantization (VQ) [14], sparse coding [1], Principal Component Analysis (PCA) [18], or Non negative matrix factorization (NMF) [37] are based on the principle that our observations can be described by a sparse subset of atoms taken from a redundant representation. These method provide a small subset of highly compressed data that is later used to guide the extraction of a target instrument.

Building a dictionary using K-SVD has been a successful approach in image processing applications [44]. However this method does not scale well to process large audio signals as the computational is unrealistic. Thus a genre specific dictionary cannot be considered in this framework.

VQ has been mainly used for audio compression, sparse coding has been used among other things for source separation [33] and PCA has been used for voice extraction [18]. However these methods do not have been used yet has a pre-processing step to build a dictionary.

Finally, in the NMF framework, some work has been done to perform a decomposition with learned dictionaries. In [15], the dictionary is built using a physical model of the instrument. This method is not adapted to build genre specific data as the model cannot easlily take into account the genre information. A second way to build the dictionary is to directly use the STFT of a target instrument signal [42]. This method does not scale well if the training data are large so it is not possible to use it to build genre specific information. Finally, another method using NMF to build a dictionary is to compute a NMF decomposition on a large training set specific to the target source [36]. After the optimization process of the NMF, the W matrix from this decomposition is used as the dictionary matrix W_{train}. This method is difficult to use on pitched instruments (i.e., harmonic instruments) and the dictionary needs to be adaptated using linear filtering og the fixed templates [19]. The fixed dictionaries provided good results for HPSS [13], however, the results has a high variance because the dictionary are learned on general data that do not take into account the large variability of drum sounds. Finally, the rank of the factorization determines the final size of the dictionary and it can be chosen small enough to obtain a strong compression of the original data. This property motivated us to build genre specific data using NMF in order to obtain compact dictionaries more appropriated the signals.

2.3 Genre information for HPSS

Current state-of-the-art unsupervised methods for HPSS such as [31] and [6] cannot be easily adapted to use genre information. In [20] the drum source separation is done using a Non-Negative Matrix Partial Co-Factorization (NMPCF). The spectrogram of the signal and the drum-only data (obtained from prior learning) are simultaneously decomposed in order to determine common basis vectors that

capture the spectral and temporal characteristics of the drum sources. The percussive part of the decomposition is constrained while the harmonic part is completely unconstrained. As a result, the harmonic part tends to decompose a lot of information from the signal and the decomposition is not satisfactory (i.e., the harmonic part contains some percussive instruments). The other major problem of this method is that it does not scale when the training data is large and the computation time is significantly larger compared to other methods.

Finally the method first introduced by [12] and detailed in [13] is a good candidate to test the genre specific dictionaries as they are easily integrated to the algorithm and do not increase the computation time. The method is detailed in Section 3.

3. STRUCTURED PROJECTIVE NMF (SPNMF)

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

3.1 Principle of the SPNMF

Using a similar model as in our preliminary work [12], let V be the magnitude spectrogram of the input data. The model is then given by

$$V \approx \tilde{V} = V_H + V_P, \tag{2}$$

with V_P the spectrogram of the percussive part and V_H the spectrogram of the harmonic part. V_H is approximated by the projective NMF decomposition [43] while V_P is decomposed by NMF components which leads to:

$$V \approx \tilde{V} = W_H W_H^T V + W_P H_P. \tag{3}$$

The data matrix is approximated by an almost orthogonal sparse part that codes the harmonic instruments $V_H = W_H W_H^T V$ and a non constrained NMF part that codes the percussive instruments $V_P = W_P H_P$. As a fully unsupervised SPNMF model does not allow for a satisfying harmonic/percussive source separation [12], we propose here to use a fixed genre specific drum dictionary W_P in the percussive part of the SPNMF.

3.2 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. A discussion about the possible use of other divergences can be found in [13].

The SPNMF model gives the optimization problem :

$$\min_{W_H, W_P, H_P \geq 0} D(V|W_H W_H^T V + W_P H_P) \tag{4}$$

A solution to this problem can be obtained by iterative multiplicative update rules following the same strategy as in [24, 43] which consists in splitting the gradient with respect to (wrt) one variable (here W_H for example) $\nabla_{W_H} D(V|\tilde{V})$ in its positive $[\nabla_{W_H} D(V|\tilde{V})]^+$ and negative parts $[\nabla_{W_H} D(V|\tilde{V})]^-$, if $\otimes$ is the Hadamard product or element-wise product. Using formula from Annex 7, the following algorithm 1 gives the SPNMF optimization process.

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 $X_H = W_H W_H^T V$

Algorithm 1: SPNMF with a fixed trained drum dictionary matrix.

3.3 Signal reconstruction

The percussive signal $x_p(t)$ is synthesized using the magnitude percussive spectrogram $X_P = W_P H_P$. To reconstruct the phase of the percussive part, we use a Wiener filter [27] to create a percussive mask as:

$$\mathcal{M}_P = \frac{X_P^2}{X_M^2 + X_P^2} \tag{5}$$

To retrieve the percussive signal as,

$$x_p(t) = InverseSTFT(\mathcal{M}_P \otimes X). \tag{6}$$

Where X is the complex spectrogram of the mixture. We use a similar process for the harmonic part.

4. CONSTRUCTION OF THE DICTIONARY

In this Section, we present in Section 4.1 the test conducted on the SiSEC 2010 database [3] in order to find the optimal size to build the genre specific dictionaries. In Section 4.2 we describe the training and the evaluation database. Finally, in Section 4.3, we detail the protocol to build the genre specific dictionaries.

4.1 Optimal size for the dictionary

The first step to build a NMF drum dictionary is to select the rank of factorization. We run the optimization tests on the public SiSec database from [3] to avoid overtraining. The dataset is composed of 4 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. Following the same protocol as in [6], we will not consider the vocal part and we will build the mixture signals from the percussive and harmonic instruments only. The signals are sampled at 44.1 kHz. We compute the STFT with a 2048 sample long Hann window with a 50% overlap. Furtheremore, the rank

of the factorization of the harmonic part for the SPNMF algortihm is $k = 100$ as in [13].

The drum signal used for the training comes from the database ENST-Drums [30] and the signal is around 10 min long. We used 30 files were the drummer is playing a *drum phrase*. We compute an NMF decomposition with different rank of factorization ($k = 12$, $k = 50$, $k = 100$, $k = 200$, $k = 300$, $k = 500$, $k = 1000$ and $k = 2000$) on the drum signal alone to obtain 8 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 compared by means of the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Artifact Ratio (SAR) of each of the separated sources using the BSS Eval toolbox provided in [41].

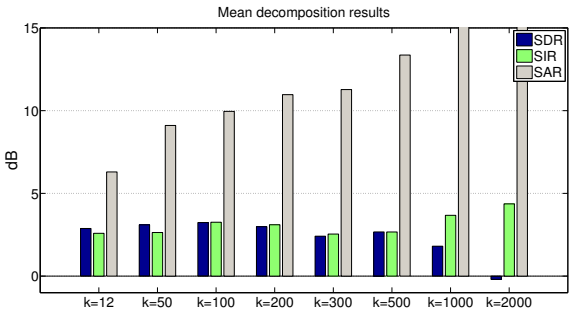

Figure 1: Influence of k on the S(D/I/A)R on the SiSEC database.

The results on figure 1 show that the optimal value for the SDR and SIR is reached for $k = 100$, then the SDR decreases for $k \geqslant 200$. The high values of SAR (for $k \geqslant 500$) are explained by the fact that the separation process is not satisfying. The harmonic signal provided by the algorithm is composed of most of the original signal therefore the SAR is very high but the decomposition quality is poor.

The optimal dictionary size for the SPNMF algorithm is $k = 100$ so we will use it in the next Section. $k = 50$ and $k = 200$ does not lead to significantly different. The method is robust and the to the rank of factorization.

4.2 Training and evaluation database

The dataset Medley-dB [5] is used for our tests. It is composed of polyphonic real-world music excerpts. It has 122 music signals and 85 of them contain percussive instruments, harmonic instruments and vocals. The signals that do not contain a percussive part are not part of the evaluation. The genres are, *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 (see Section 2), 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, the *Pop* label (for example) are similar to the *Singer/Songwriter*.

Genre	Artist Song
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.

The training dataset is built using 3 songs of each genre. The songs used for the training part are not included in the evaluation. To compare the genre specific dictionaries, we build a non specific/universal dictionary using half of one song of each genre. Finally, a last dictionary is built using around 10 min of pure drum signals from the ENST-Drums database as in Section 4.1. The files from medley-dB selected for the training are given in Table 1.

4.3 Genre specific dictionaries

The NMF model is given by (1). 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. Here we build genre specific drum dictionaries using the medley-dB database.

With the results from Section 4.1 the dictionaries are built as follows. For every genre specific subset of the training database, we perform a NMF on the drum signals with $k = 100$. Then the W matrices of the NMF are used in the SPNMF algorithm as the matrix W_P (see algorithm 1).

5. RESULTS

In this Section, we present the results of the SPNMF with the genre specific dictionaries on the evaluation database from Medley-dB.

5.1 Comparison of the dictionaries

In this section we present the results of the SPNMF algorithm with the genre specific dictionaries on the 64 remaining song from test database Medley-dB. We perform an HPSS on the audio files using the SPNMF with the 9 dictionaries created in Section 4.3. The results on each of the songs are then sorted by genres and the average results are displayed using box-plots. 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 Figures 2, 3 and 4 show the SDR, SAR and SIR results for all the dictionaries on the *Pop* subsection. It gives us an overall idea of the performance of the dictionaries on the same database. The *Pop* dictionary leads to the highest SDR and SIR. The non specific dictionaries are not performing as well as the *Pop* dictionary. On this database, the genre specific data gives relevant information to the algorithm. As stated in Section 4.2, some genres are similar to other. This explains why the *Rock* and the *Singer* dictionaries are providing good results too. An interesting result is that compared to the non specific dictionary, the *Pop* dictionary has a lower variance. Genre information allows for a higher robustness to the variety of the songs within the same genre.

The dictionary built on the ENST-drums is giving very similar results to the universal dictionary built on the Medley-dB database. For the sake of concision we only display the results using the universal dictionary from Medley-dB.

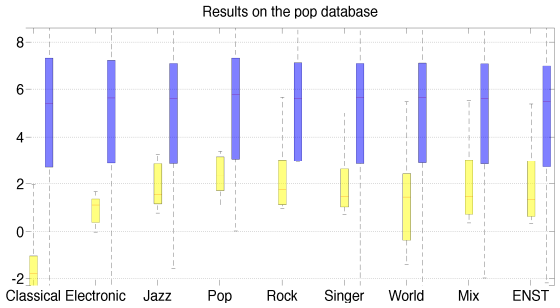

Figure 2: Percussive (left bar)/Harmonic (right bar) SDR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

On Table 2, 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 on 5 of the 7 genres. The results are discussed in the next Section.

5.2 Discussion

On the database *Singer/Songwriter*, *Pop*, *Rock*, *Jazz* and *World/Folk*, the genre specific dictionaries outperform the universal dictionary.

The information from the music of the same genre is not altered by the NMF compression and the drum templates are closer to the target drum. The databases *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 thus the drum dictionary does not contain a sufficient amount of information to outperform the universal dictionary. Because of these two factors, the genre specific dictionaries are not performing correctly.

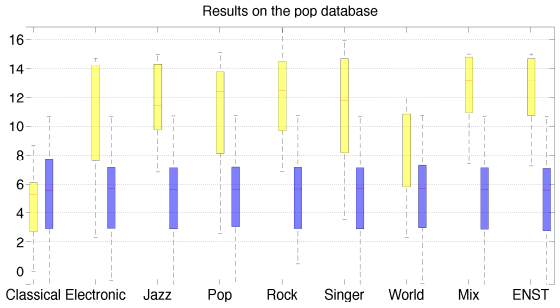

Figure 3: Percussive (left bar)/Harmonic (right bar) SIR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

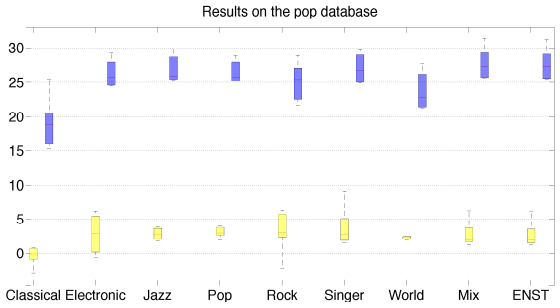

Figure 4: Percussive (left bar)/Harmonic (right bar) SAR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

Overall the harmonic separation is giving much better results than the percussive extraction. The fixed dictionaries are creating artefact as the percussive templates do not correspond exactly to the target drum signal. A possible way to alleviate this problem is to adapt the dictionaries but this requires the use of hyper parameters and that is not the philosophy of this work [12].

6. CONCLUSION

Using genre specific information in order to build more relevant drum dictionaries is a powerful method to improve the HPSS. The dictionaries still have an imprint of the genre after the NMF decomposition and the additional information is properly used by the SPNMF to improve the source separation quality. 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 building a blind method

Genre	Classical	Electronic/Fusion	Jazz	Pop	Rock	Singer/Songwriter	World/Folk
Percussive separation							
Genre specific (dB)							
SDR	-1.64	-0.63	**0.36**	**2.45**	**-0.17**	**0.64**	**0.42**
SIR	8.21	15.17	**9.60**	12.34	**19.79**	11.45	**6.08**
SAR	5.88	0.31	**2.08**	**3.36**	0.31	**4.46**	**16.29**
Non specific (dB)							
SDR	**-0.04**	**-0.25**	-0.68	2.01	-2.15	-0.01	-3.57
SIR	**11.3**	**17.01**	9.57	**12.60**	18.30	**13.04**	2.82
SAR	**8.07**	**0.39**	0.87	2.74	**2.34**	1.83	12.08
Harmonic Separation							
Genre specific (dB)							
SDR	**7.49**	**1.63**	**13.05**	**5.06**	**2.14**	7.20	**4.86**
SIR	**10.60**	**1.84**	**13.27**	**5.02**	2.19	**11.45**	**13.50**
SAR	18.19	23.48	28.50	24.48	**35.97**	28.48	**22.65**
Non specific (dB)							
SDR	6.04	1.33	12.71	4.78	1.92	**7.46**	4.64
SIR	7.14	1.36	12.82	4.85	**2.86**	7.50	13.27
SAR	**27.21**	**27.72**	**29.87**	**26.17**	34.25	**31.87**	21.64

Table 2: Average SDR, SIR and SAR results on the Medley-dB database.

to select the genre specific dictionary in order to perform the same technique on database where the genre information is not available.

7. SPNMF WITH THE IS 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 $I \in \mathbb{R}^{f\times t}; \forall i,j \quad I_{i,j} = 1$.

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

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

8. REFERENCES

[1] S.A. Abdallah and M.D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *IEEE Transactions on Neural Networks*, pages 179–196, 2006.

[2] M. Aharon, M. Elad, and Alfred A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, pages 4311–4322, 2006.

[3] 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.

[4] J.J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, pages 83–93, 2003.

[5] 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.

[6] 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*, pages 1–17, 2014.

[7] D. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, pages 51–60, 2007.

[8] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *Proc. of IEEE ICASSP*, pages 753–756, 2000.

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

[10] D. Fitzgerald. Harmonic/percussive separation using median filtering. In *Proc. of DAFx*, 2010.

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

[12] Hidden for the review process.

[13] Hidden for the review process.

[14] A. Gersho and R.M. Gray. *Vector quantization and signal compression*. Springer Science & Business Media, 2012.

[15] R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. of IEEE ICASSP*, 2011.

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

[17] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang, and Ke Hu. A tandem algorithm for singing pitch extraction and voice separation from music accompaniment. *Transactions on Audio, Speech, and Language Processing.*, 20(5):1482–1491, 2012.

[18] P. Huang, S.D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *Proc. of IEEE ICASSP*, pages 57–60, 2012.

[19] 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.

[20] Minje Kim, Jiho Yoo, Kyeongok Kang, and Seungjin Choi. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *Journal of Selected Topics in Signal Processing*, pages 1192–1204, 2011.

[21] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech, and Language Processing.*, pages 255–266, 2008.

[22] A. Lampropoulos, P. Lampropoulou, and G. Tsihrintzis. Musical genre classification enhanced by improved source separation technique. In *Proc. of ISMIR*, pages 576–581, 2005.

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

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

[25] 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*, pages 291–301, 2008.

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

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

[28] C. McKay and I. Fujinaga. Musical genre classification: Is it worth pursuing and how can it be improved? In *Proc. of ISMIR*, pages 101–106, 2006.

[29] 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.

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

[31] 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 *Proc. of EUSIPCO*, 2008.

[32] J. Paulus and T. Virtanen. Drum transcription with non-negative spectrogram factorisation. In *Proc. of EUSIPCO*, pages 1–4, 2005.

[33] M.D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies. Sparse representations in audio and music: from coding to source separation. *Proceedings of the IEEE*, pages 995–1005, 2010.

[34] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagayama. Autoregressive mfcc models for genre classification improved by harmonic-percussion separation. In *Proc. of ISMIR*, pages 87–92, 2010.

[35] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *Transactions on Audio, Speech, and Language Processing.*, pages 1759–1770, 2012.

[36] M.N Schmidt and R.K. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Proc. of INTERSPEECH*, 2006.

[37] P. Smaragdis and JC. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003.

[38] I. Tošić and P. Frossard. Dictionary learning. *IEEE Transactions on Signal Processing*, pages 27–38, 2011.

[39] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, pages 293–302, 2002.

[40] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing.*, pages 528–537, 2010.

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

[42] C. Wu and A. Lerch. Drum transcription using partially fixed non-negative matrix factorization. In *Proc. of EUSIPCO*, 2008.

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

[44] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *Proc. of IEEE CVPR*, pages 2691–2698. IEEE, 2010.