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 use apriori information like training data and dictionary to achieve a decomposition. Audio signals are diverse and it is generally impossible to build prior information that is relevant on any audio signal. Most of the method are tested on small databases that do not allow exhaustive comparison of the algorithms. When using a large database, the algorithms are not robust to the wide variety of audio signals and they do not provide satisfying results without user intervention. In this article, we focus on the particular case of harmonic/percussive source separation and we propose to use musical genre information to guide the decomposition on a large database. Our method proves that the musical genre is a relevant feature as the dictionaries built using genre specific information perform better that an universal drum dictionary.

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 : up-mixing [12](spatialization of the sources) or automatic transcription [9] (it is easier to work on single source). The 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 obtain a true evaluation rather than particular case results.

The large variety of audio signals can be classified into different musical genres [38]. 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 have been used to improve chord transcription [26, 30] or downbeat detection [15] algorithms. Genre information is obtained using annotated labels whom when the genre information is not available, can be retrived using automatic genre classification [29, 38].

In the context of BASS, 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

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

with $W \in \mathbb{R}_+^{n\times k}$, $H \in \mathbb{R}_+^{k\times m}$ and where k is the rank of factorization [24]. 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 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 [10] can be used to initialize the learning process. The downside of these approachs are 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} is learned from a big database in order to separate an instrument [18, 40]. These methods require minimum tuning from the user. However, the dictionaries must match the target instruments for satisfying performances.

In this paper, we focus on the task of harmonic/percussive source separation (HPSS) using the method developed in [23]. We adapt the algorithm to use a trained drum dictionary to extract the percussive instruments. This method is explained in detail in [22]. The problem of using a fixed dictionary matrix is that within a database, the same instrument can sound differently depending on the recording conditions and post processing treatments. In order to represent correctly one instrument, one can decide to learn a dictionary on a large database but take the risk of over-fitting the data. In order to avoid this problem and to build effective dictionaries, we decided to use genre specific training data. As songs with identical style 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 NMF drum dictionaries that obtain consis-

tent and robust results on a HPSS task. The genre specific dictionaries are able to improve the separation score compared to a universal dictionary.

Section 2 defines the context of our work, Section 3 presents the algorithm used while Section 4 describes the construction of dictionaries specific. Finally Section 5 details the results of the HPSS on 65 audio files and 6 conclude our work.

2. TOWARD GENRE SPECIFIC INFORMATION

2.1 Genre information

Musical genre is one of the most prominent high level music descriptor. Electronic Music Distribution have become more popular in recent years and music catalogues never stop to increase (the biggest online services 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. If an explicit definitions of musical genres is still out of reach [4], musical genre classification can be performed automatically using different set of features [27].

Source separation have been used extensively in order to help the classification process [21, 35]. However, genre information have not been 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...) and each version is usually densely sampled [37]. 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) [13], Principal Component Analysis (PCA) [17], sparse coding [1] or factorisations non-negative matrix (NMF) cite Smaragdis2003 are based on the principle that our observations can be described by a sparse subset of atoms taken from a redundant representation hence the idea to use these methods to create genre specific drum dictionaries.

Building using dictionary using K-SVD is a successful method for image processing [42], however, this method do not scale well to process large audio signal and a genre specific dictionary cannot be considered with this method.

VQ have been mainly used for audio compression, sparse coding have been used amoung other things for source separation [34], and PCA have been used for voice extraction. However these methods do not have been used yet has a pre-processing step to build a dictionary.

Finaly, in the NMF framework, some work has been done to perform a decomposition with dictionaries. In [14], the dictionary is built using a physical model of the instrument. This method is not adapted to build genre specific data. Another method using NMF to build a dictionary is to compute a NMF decomposition on a large training set specific to the target source. After the optimization process of the NMF, the W matrix from the 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 [18], however, it provided good results for HPSS [22]. The rank of factorization determines the final size of the dictionary and it can be chosen small to obtain a strong compression of the original data. This property motivated us to build genre specific data using NMF. The selection of the rank of factorization will be discussed in Section 4.1.

2.3 Genre information for HPSS

Harmonic/percussive source separation has numerous applications as a preprocessing step. For example most multi-pitch estimation models [20], instruments recognition [8] and melody extraction [36] algorithms are much more efficient on harmonic data only. Similarly, beat tracking [7] and drum transcription algorithms [33] are more accurate if the harmonic instruments are not part of the signal. Finally, using the HPSS algorithm [11] as a preprocessing step increases the performance for singing pitch extraction and voice separation [16].

Current state-of-the-art unsupervised methods for HPSS such as [32] and [6] cannot be easily adapted to use genre information. However, the supervised algorithm however can use genre specific dictionary. In [19] 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 drum sources. The percussive part of the decomposition is constrained while the harmonic part is completely unconstrained. As a result, it tends to decompose a lot of information (i.e., the harmonic part contains some percussive instruments) from the signal and the decomposition is not satisfactory. The other major problem of this method is that it does not scale when the training data are large and the computation time in unrealistic compared to other methods.

Finally the method first introduced by [23] and detailed in [22] is a good candidate to test the genre specific dictionaries as they are easily integrated to the method 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 [23], 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 [41] while V_P is decomposed by NMF components as :

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

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 $V_P = W_P H_P$. We use here a fixed genre specific 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 [23].

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.

The SPNMF model gives the cost function :

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

A solution of this problem can be obtained by iterative multiplicative update rules following the same strategy as in [25, 41] 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 $[\nabla_{W_H} D(V|\tilde{V})]^-$. If $\otimes$ is the Hadamard product or element-wise product. The algorithm 1 gives us 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 the 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 generalized Wiener filter [28] to create a percussive mask as:

$$\mathcal{M}_P = \frac{X_P^2}{X_M^2 + X_P^2}. \quad (5)$$

To retrieve the percussive signal as,

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

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

$$\mathcal{M}_H = \frac{X_H^2}{X_M^2 + X_P^2}, \quad (7)$$

and:

$$x_h(t) = InverseSTFT(\mathcal{M}_H \otimes X). \quad (8)$$

4. CONSTRUCTION OF THE DICTIONARY

In this section, we present in Section 4.1 the test conducted on the SiSec database in order to find the optimal parameter 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 [6], we will not consider the vocal part and we will build the mixture signals from the percussive and harmonic instruments only. All the signals are sampled at $44.1kHz$. We compute the STFT with a and 2048 sample-long Hann window with a 50% overlap. Furtheremore, the rank of factorization of the harmonic part for the SPNMF algortihm is $k = 100$ as in [22].

The drum signal used for the training comes from the database ENST-Drums [31] and the signal is around 10 min long. We used 30 files from the database were the drummer is playing a *drum phrase*. We compute an NMF decomposition with different rank of factorization ($k = 12$, $k = 50$, $k = 100$, $k = 200$, $k = 300$, $k = 500$, $k = 1000$ 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 [39].

The results on the figure 1 show that the optimal value for the SDR and SIR is reached for $k = 100$, then the SDR decrease for $k \geqslant 200$. 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 the SAR is very high but the decomposition quality is poor.

The optimal dictionary size for the SPNMF algorithm is $k = 100$ and that what we will in the next Section. The test using $k = 50$ and $k = 300$ are not significantly different,

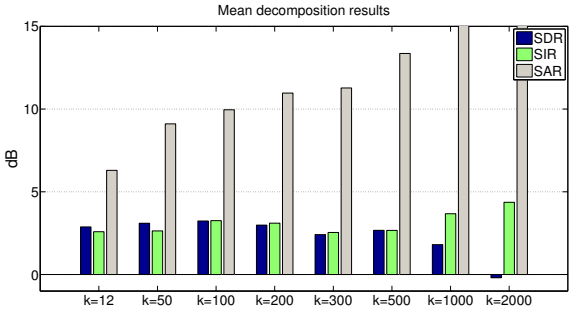

Figure 1: Average results on the SiSec database.

it means that 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. We are 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 (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*.

The training dataset is built using 3 song of each genre. 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 half of one song of each genre. Finally, a last dictionary is built using around 10min of pure drum signals from the ENST-Drums database similarly 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 dictionary using the medley-dB database. Using dictionary specific to the genre of music allows us to have dictionaries that a more specific to the signal to decompose.

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

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.

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 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-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 Figures 2, 3 and 4 show 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 and SIR results. The non specific dictionary is not performing as well as the *Pop* dictionary. On this database, the genre specific method is giving relevant information to the algorithm. As stated in Section 4.2, some genre are similar to other. This explains why the *Rock* and the *Singer* dictionaries are giving 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.

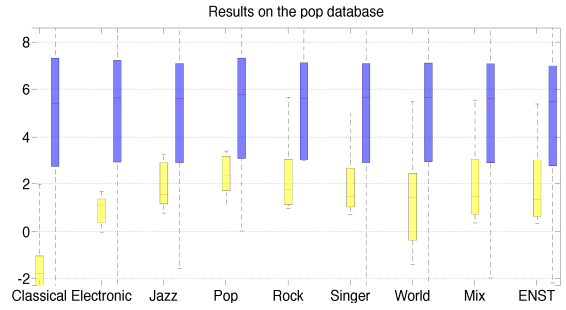

Figure 2: Percussive (left bar)/Harmonic (right bar) SDR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

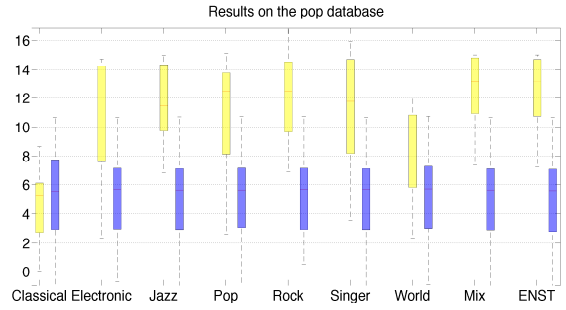

Figure 3: Percussive (left bar)/Harmonic (right bar) SIR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

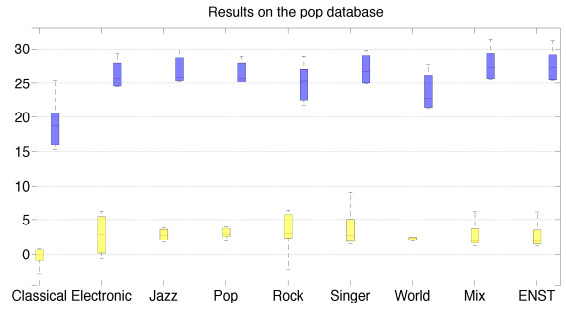

Figure 4: Percussive (left bar)/Harmonic (right bar) SAR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

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 correlation between the music of the same genre is not altered by the NMF compression and the drum templates are more suited to the decomposition. 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 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.

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.

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 even after the NMF decomposition and the additional information is properly used by the SPNMF to improve the source separation. 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 to select the genre specific dictionary in order to perform the same technique on database where the genre information.

7. ANNEXE

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

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: Mean harmonic SDR, SIR and SAR results on the Medley-dB database.

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] 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.

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

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

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] C. Laroche, M. Kowalski, H. Papadopoulos, and G. Richard. Spnmf with drum dictionaries for harmonic/percussive source separation. *submitted to IEEE Transactions on Acoustics, Speech and Signal Processing*.

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

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

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

[26] 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.

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

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

[29] 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.

[30] 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.

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

[32] 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.

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

[34] 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.

[35] 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.

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

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

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

[39] 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.

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

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

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