Genre specific dictionaries for harmonic/percussive source separation

Clément Laroche, Hélène Papadopoulos, Matthieu Kowalski, Gaël Richard

GENRE SPECIFIC DICTIONARIES FOR HARMONIC/PERCUSSIVE SOURCE SEPARATION

Clément Laroche[1,2] **Hélène Papadopoulos**[2] **Matthieu Kowalski**[2,3] **Gaël Richard**[1]

[1] LTCI, CNRS, Télécom ParisTech, Univ Paris-Saclay, Paris, France
[2] Univ Paris-Sud-CNRS-CentraleSupelec, L2S, Gif-sur-Yvette, France
[3] Parietal project-team, INRIA, CEA-Saclay, France
[1]name.lastname@telecom-paristech.fr, [2]name.lastname@lss.supelec.fr

ABSTRACT

Blind source separation usually obtains limited performance on real and polyphonic music signals. To overcome these limitations, it is common to rely on prior knowledge under the form of side information as in *Informed Source Separation* or on machine learning paradigms applied on a training database. In the context of source separation based on factorization models such as the *Non-negative Matrix Factorization*, this supervision can be introduced by learning specific dictionaries. However, due to the large diversity of musical signals it is not easy to build sufficiently compact and precise dictionaries that will well characterize the large array of audio sources. In this paper, we argue that it is relevant to construct genre-specific dictionaries. Indeed, we show on a task of harmonic/percussive source separation that the dictionaries built on genre-specific training subsets yield better performances than cross-genre dictionaries.

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 [9] (spatialization of the sources) or automatic transcription [35] (it is easier to work on single sources). The separation task is difficult due to the complexity and the variability of the music mixtures.

The large collection of audio signals can be classified into various musical genres [34]. Genres are labels created and used by humans for categorizing and describing music. They have no strict definitions and boundaries but particular genres share 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 [23, 27] or downbeat detection [13] 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 [26, 34]. Such classification have never been used to guide a source separation problem and this may be due to the lack of annotated databases. The recent availability of large evaluation databases for source separation that integrate genre information motivates the development of such approaches. Furthermore, 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 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 [21]. In audio signal processing, the input data is usually a Time-Frequency 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 [7]. 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 database in order to separate the target instrument [16, 37]. Such method requires minimum tuning from the user. However, within different music pieces of an evaluation database, the same instrument can sound differently depending on the recording conditions and post processing treatments.

In this paper, we focus on the task of Harmonic Percussive Source Separation (HPSS). HPSS has numerous applications as a preprocessing step for other audio tasks. For example the HPSS algorithm [8] can be used as a preprocessing step to increase the performance for singing pitch extraction and voice separation [14]. Similarly, beat tracking [6] and drum transcription algorithms [29] are more accurate if the harmonic instruments are not part of the analyzed signal.

We built our algorithm using the method developed in [20]: an unconstrained NMF decomposes the audio signal in a sparse orthogonal part that are well suited for representing the harmonic component, while the percussive part is represented by a regular nonnegative matrix factorization decomposition. In [19], we have adapted the algorithm using a trained drum dictionary to improve the extraction of the percussive instruments. As the user databases typically cover a wide variety of genres, instrumentation may strongly differ from one piece to another. In order to better manage the variability and to build effective dictionaries, we propose here 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 gives 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 (i.e. a cross-genre dictionary).

The rest of the paper is organized as follows. 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 30 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 [3], musical genre classification can be performed automatically [24].

Source separation has been used extensively in order to help the genre classification process [18,30] but, at the best of our knowledge, the genre information has never been exploited to guide source separation algorithm.

2.2 Methods for dictionary learning

Audio data is largely redundant as it often contains multiple correlated versions of the same physical event (note, drum hits...) [33] 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) [1], Vector Quantization (VQ) [10], Principal Component Analysis (PCA) [15], or Non negative matrix factorization (NMF) [32] are based on the principle that our observations can be described by a sparse subset of atoms taken from a redundant representation. These methods provide a small subset of relevant templates that are later used to guide the extraction of a target instrument.

Building a dictionary using K-SVD has been a successful approach in image processing [39]. However this method does not scale well to process large audio signals as the computational time is unrealistic. Thus a genre specific dictionary scenario cannot be considered in this framework.

VQ has been mainly used for audio compression [10] and PCA has been used for voice extraction [15]. However these methods have not been used yet as 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 [12], a dictionary is built using a physical model of the piano. This method is not adapted to build genre specific dictionaries as the model cannot easily take into account the genre information. A second way to build a dictionary is to directly use the STFT of an instrument signal [37]. This method does not scale well if the training data is large, thus it is not possible to use it to build genre specific dictionaries. Finally, another method to build a dictionary is to compute a NMF decomposition on a large training set specific to the target source [31]. After the optimization process of the NMF, the W matrix from this decomposition is used as a fixed dictionary matrix W_{train}. This method does not give satisfying results on pitched instruments (i.e., harmonic instruments) and the dictionary needs to be adaptated for example using linear filtering on the fixed templates [16]. Compared to state of the art methods, fixed dictionaries provide good results for HPSS [19]. However, the results have a high variance because the dictionaries are learned on general data that do not take into account the large variability of drum sounds. A nice property of the NMF framework is that 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. The limitations of the current methods motivated us to build genre specific data using NMF in order to obtain relevant compact dictionaries.

2.3 Genre information for HPSS

Current state-of-the-art unsupervised methods for HPSS such as complementary diffusion [28] and constrained NMF [5] cannot be easily adapted to use genre information. We will not discuss these methods in this article.

However supervised methods can be modified to utilize genre information. In [17] the drum source sepa-

ration 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 separation is not satisfactory (i.e., the harmonic part contains some percussive instruments). A drawback 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.

By contrast, the approach introduced and detailled in [19, 20] appears to be a good candidate to test the genre specific dictionaries: they can be easily integrated to the algorithm without increasing the computation time.

3. STRUCTURED PROJECTIVE NMF (SPNMF)

3.1 Principle of the SPNMF

Using a similar model as in our preliminary work [20], 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 [38] 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 [20], 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 [19].

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 [22, 38]. Using formula from Appendix 7, the optimization process is given in Algorithm 1, where $\otimes$ is the Hadamard product and all division are element-wise operation.

Input: $V \in \mathbb{R}_+^{m \times n}$ and $W_{train} \in \mathbb{R}_+^{m \times e}$ Output: $W_H \in \mathbb{R}_+^{m \times k}$ and $H_P \in \mathbb{R}_+^{e \times n}$ 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 [25] to create a percussive mask as:

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

To retrieve the percussive signal as:

$$x_p(t) = \text{SFTF}^{-1}(\mathcal{M}_P \otimes X). \tag{6}$$

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

4. CONSTRUCTION OF THE DICTIONARY

In this section we detail the building process of the drum dictionary. We present in Section 4.1 tests conducted on the SiSEC 2010 database [2] 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 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. The first step to build a NMF drum dictionary is to select the rank of factorization. In order to avoid overfitting, the algorithm is optimized using databases different from the database used for evaluation, described in Section 4.2.

We run the optimization tests on the public SiSec database [2]. The database is composed of four polyphonic real-world music excerpts and each music signal contains percussive, harmonic instruments and vocals. The duration of the recordings is ranging from 14 to 24 s. In the context of HPSS, following the same protocol as in [5], we do not consider the vocal part and we 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. Furthermore, the rank of factorization of the harmonic part of the SPNMF algorithm is set to $k = 100$, as in [19].

A fixed drum dictionary is built using the database ENST-Drums [11]. For this, we concatenate 30 files where the drummer is playing a *drum phrase* that result in an excerpt of around 10 min duration. We then compute an NMF decomposition with different ranks 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.

These dictionaries are then used to perform a HPSS on the four songs of the SiSEC database using the SPNMF algorithm (see Algorithm 1). 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 [36].

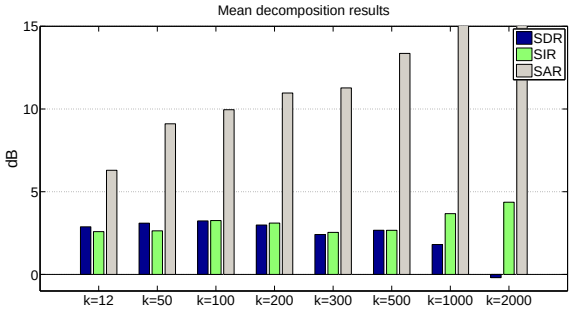

Figure 1: Influence of k on the S(D/I/A)R on the SiSEC database.

The results in Figure 1 show that the optimal value for the SDR and SIR is reached for $k = 100$, then the SDR decreases for $k \geqslant 200$. For $k \geqslant 500$ the harmonic signal provided by the algorithm contains most of the original signal therefore the SAR is very high but the decomposition quality is poor. For the rest of the article, the size of the drum dictionaries will be $k = 100$.

4.2 Training and evaluation database

The evaluation tests are conducted on the Medley-dB database [4] composed of polyphonic real-world music excerpts. It consists in 122 music signals and 85 of them contain percussive instruments, harmonic instruments and vocals. The signals that do not contain a percussive part are excluded from evaluation. The genres are distributed as follows: *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). It is important to note that, because the notion of genre is quite subjective (see Section 2), the Medley-dB database uses general genre labels that cannot be considered to be precise. 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. Moreover, as we are only working with the instrumental part of the song (the vocals are omitted), the *Pop* label (for example) is similar to the *Singer/Songwriter*. We separate the database into training and evaluation files, as detailed in the next section.

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.

4.3 Genre specific dictionaries

Seven genre-specific drum dictionaries are built using 3 songs of each genre. In addition, a cross-genre drum dictionary is built using half of one song of each genre. Finally, a dictionary is built using the 10 min excerpt of pure drum signals from the ENST-Drums database described in Section 4.1. The Medley-dB files selected for training are given in Table 1 and excluded from evaluation.

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 $k = 100$. The resulting W matrices of the NMF are then used in the SPNMF algorithm as the W_P matrix (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

We perform a HPSS on the audio files using the SPNMF algorithm with the 9 dictionaries built in Section 4.3. The results on each song 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.

Figures 2, 3 and 4 show the SDR, SAR and SIR results for all the dictionaries on the *Pop* subset, giving an overall idea of the performance of the dictionaries inside a specific sub-database. The *Pop* dictionary leads to the highest SDR and SIR and the non specific dictionaries are not performing as well. On this sub-database, the genre specific data gives relevant information to the algorithm. As stated in Section 4.2, some genres are similar to others, explaining why the *Rock* and the *Singer* dictionaries are also providing good results. An interesting result is that compared to the non specific dictionaries, the *Pop* dictionary has a lower variance. Genre information allows for a higher robustness to the variety of the songs within the same genre. Samples of the audio results can be found on the website [1].

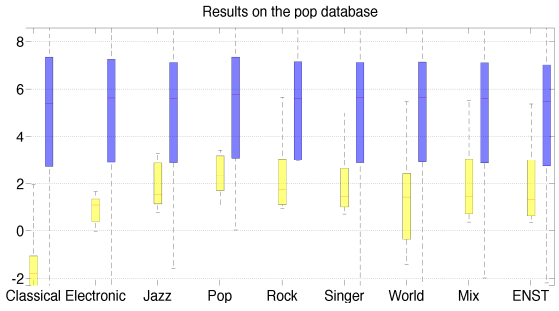

Figure 2: Percussive (left bar)/Harmonic (right bar) SDR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

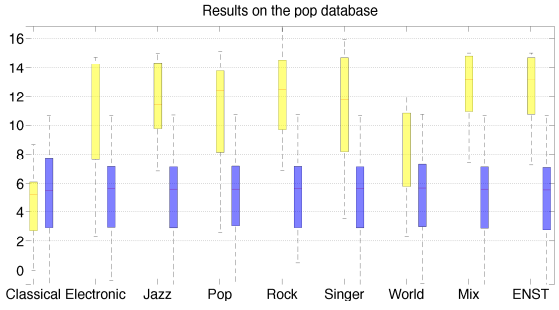

Figure 3: Percussive (left bar)/Harmonic (right bar) SIR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

On Table 2, we display the mean separation score for all the genre specific dictionaries compared to the non specific dictionary. The dictionary built on the ENST-drums is giving results very similar 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. On the database *Singer/Songwriter*, *Pop*, *Rock*, *Jazz* and *World/Folk*, the genre specific dictionaries outperform the universal dictionary on the harmonic and percussive separation.

[1] https://goo.gl/4X2jk5

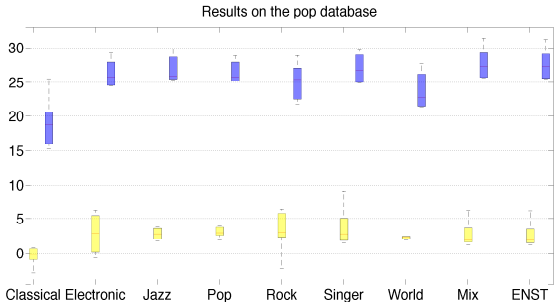

Figure 4: Percussive (left bar)/Harmonic (right bar) SAR results on the *Pop* sub-database using the SPNMF with the 9 dictionaries.

5.2 Discussion

The cross-genre dictionary as well as the ENST-drum dictionary are outperformed by the genre specific dictionaries. The information from the music of the same genre is not altered by the NMF compression and provides drum templates 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. As a result, in both cases 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.

It can be noticed that 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 would be to adapt the dictionaries but this would require the use of hyper parameters and that is not the philosophy of this work [20].

6. CONCLUSION

Using genre specific information in order to build more relevant drum dictionaries is a powerful approach 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 separating 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 is not available.

Genre	Classical	Electronic/Fusion	Jazz	Pop	Rock	Singer/Songwriter	World/Folk
Percussive separation							
Genre specific (dB)							
SDR	-1.6	-0.6	**0.4**	**2.5**	**-0.2**	**0.6**	**0.4**
SIR	8.2	15.2	**9.6**	12.3	**19.8**	11.5	**6.1**
SAR	5.9	0.3	**2.1**	**3.4**	0.3	**4.5**	**16.3**
Non specific (dB)							
SDR	**-0.0**	**-0.3**	-0.7	2.0	-2.2	-0.0	-3.6
SIR	**11.3**	**17.0**	9.6	**12.6**	18.3	**13.0**	2.8
SAR	**8.1**	**0.4**	0.9	2.7	**2.3**	1.8	12.1
Harmonic Separation							
Genre specific (dB)							
SDR	**7.5**	**1.6**	**13.0**	**5.1**	**2.1**	7.2	**4.9**
SIR	**10.6**	**1.8**	**13.3**	**5.0**	2.2	**11.5**	**13.5**
SAR	18.2	23.5	28.5	24.5	**36.0**	28.5	**22.7**
Non specific (dB)							
SDR	6.0	1.3	12.7	4.8	1.9	**7.5**	4.6
SIR	7.1	1.4	12.8	4.9	**2.9**	7.5	13.3
SAR	**27.2**	**27.7**	**29.9**	**26.2**	34.3	**31.9**	21.6

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

7. APPENDIX: SPNMF WITH THE IS DIVERGENCE

The Itakura Saito divergence gives us the problem,

$$\min_{W_H, W_P, H_P \geq 0} \frac{V}{\tilde{V}} - log(\frac{V}{\tilde{V}}) - 1.$$

The gradient wrt W_H gives

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

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

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

with

$$\phi_{i,j} = (\frac{I}{W_H W_H^T V + W_P H_P})_{i,j}.$$

and $I \in \mathbb{R}^{f \times t}; \forall i, j \quad I_{i,j} = 1$.

Similarly, the gradient wrt H_P gives

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

and

$$[\nabla_{H_P} D(V|\tilde{V})]^+ = 2W_P^T W_H W_H^T V + W_P^T W_P H_P.$$

8. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[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*, pages 1–17, 2014.

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

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

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

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

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

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

[12] 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.

[13] 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.

[14] C. Hsu, D. Wang, J.R. Jang, and K. Hu. A tandem algorithm for singing pitch extraction and voice separation from music accompaniment. *IEEE Transactions on Audio, Speech, and Language Processing.*, pages 1482–1491, 2012.

[15] 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.

[16] 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.

[17] M. Kim, J. Yoo, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *Journal of Selected Topics in Signal Processing*, pages 1192–1204, 2011.

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

[19] C. Laroche, M. Kowalski, H. Papadopoulous, and G.Richard. Structured projective non negative matrix factorization with drum dictionaries for harmonic/percussive source separation. *Submitted to IEEE Transactions on Acoustics, Speech and Signal Processing*.

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

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

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

[23] 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.

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

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

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

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

[32] 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.

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

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

[35] 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.

[36] 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.

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

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

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