PAPER TEMPLATE FOR ISMIR 2015

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

ABSTRACT

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

1. INTRODUCTION

Audio Source separation is a challenging task and fully automatic system is still out of reach, but a number of algorithm involving a human operator are starting to yield satisfactory results. Supervised algorithms use high-level musical information to improve the separation quality of the algorithm. In the context of blind source separation, Non-negative Matrix Factorization (NMF) is a widely used method for source separation. The goal of NMF is to approximate a data matrix $V \in \mathbb{R}_+^{n\times m}$ as $V \approx \tilde{V} = WH$ with $W \in \mathbb{R}_+^{n\times k}$, $H \in \mathbb{R}_+^{k\times m}$ and where k is the rank of factorization [1]. In audio signal processing, the input data is usually a Time-Frequency (TF) representation such as a short time Fourier transform (STFT) or a constant-Q transform spectrogram. Blind source separation is a difficult problem and the plain NMF decomposition does not provide satisfying results. To perform a satisfying results, 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 [2] can be used to initialize the learning process. The downside of this approach is that it requires well organized prior information that is not always available. Another supervised method consists in performing prior training on specific databases. For example a dictionary matrix W_{train} can be learnt from a big database in order to separate an instrument [3, 4]. A common method to build a dictionary for NMF is to perform a decomposition on a large training set. After the convergence, the W matrix from the decomposition is used as the dictionary matrix W_{train} in the separation [3]. Another method is detailed in [4], a dictionary matrix is created by extracting template spectra from isolated drum samples. The dictionary is then used in a NMF decomposition to perform drum transcription. This method requires minimum tuning from the user. However, the dictionary should match the target instrument for satisfying performances. The problem of recent method using dictionary matrices is that, within a database, an instrument can sound differently depending on the recording condition and post processing treatment. In order to represent correctly one instrument, ones can decide to learn a dictionary on a large database, however, the problem of overfitting the data exist. In order to overcome this problem and to be able to build effective dictionaries we decided to use genre specific training data. Genre specific information can provide an insight on the structure of the audio signal. Music from the same genre share similar chords and rhythm and the resemblance between two pieces of music have been used to perform chord transcription [5, 6] or for downbeat detection [7]. In this paper, we focus on the task of harmonic/percussive source separation (HPSS) using the method developed in [8]. We adapt the method to be used with a drum dictionary to extract the percussive instruments. This method is explained in detail in the preprint. The main contribution of this article is that we developed a genre specific method to build a drum NMF dictionary that obtains consistent results on a HPSS task. Overall using a fixed dictionary for drum extraction is an underused method in the literature as it is difficult to create a drum dictionary that provide robust results on a large variety of signal. By using genre specific dictionary we were able to improve the separation score and decrease the computation time as the dictionary are smaller in size.

2. PRESENTATION OF THE SPNMF

2.1 Overview

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

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

where $||.||^2$ is the squared Euclidean distance.

The ONMF [?] consists in solving the following problem:

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

In this method, orthogonality between nonnegative basis functions is enforced during the optimization process. In practice, it seems that PNMF and ONMF lead to similar decompositions, as the W matrix estimated by PNMF is almost orthogonal (i.e., $\|W^TW - I_k\|^2$ is small). The links between PNMF, ONMF and the regular NMF are discussed in the next section.

2.2 On the equivalence between NMF, PNMF and ONMF

Using a squared Euclidean distance between the data matrix V and its approximation WH, the NMF problems reads:

$$\min_{W,H\geq 0} \|V - WH\|^2 ,$$

where PNMF (resp. ONMF) adds the constraint $H = W^TV$ (resp. $W^TW = I$). Let us assume that V admits an NMF decomposition without any errors, i.e., one can find W and H of rank k such that $V = WH$. Then, one can easily prove that necessarily $H = (W^TW)^{-1}W^T$. Now, as demonstrated in [**?**], an invertible matrix is nonnegative if and only if it is a monomial matrix, that is, up to a scaling and permutation matrix, we necessarily have $W^TW = I$. This result is summarized in the following theorem:

In practice, the assumption $V = WH$ does not hold as soon as $k < \min(n, m)$, hence the motivation to introduce PNMF and ONMF. However, it is interesting to stress that the orthogonality of W is a requirement to obtain a true projector for H. Moreover, in practice, W_{pnmf} is "almost" orthogonal, leading to an "almost projection" for H_{pnmf} as motivated by the authors in [?]. This remark has encouraged us to build upon PNMF instead of ONMF in the next section.

2.3 Principle

As stated in [**?**], harmonic instruments have sparse basis functions whereas percussive instruments have much flatter spectra. As the columns of W are orthogonal, when two sources overlap in the Time-Frequency (TF) domain only one basis function will represent the mixture which is not adequate for efficient separation. To overcome this problem, we propose to add a standard NMF decomposition term to the PNMF. With a similar technique as in [**?**], we increase the rank of the PNMF. Let $k = k' + e$ with e being the number of additional components. We can expect that most of the harmonic components will be represented by the orthogonal part while the percussive ones will be in the regular NMF components. Let V be the magnitude spectrogram of the input data. The model is then given by

$$V \approx \tilde{V} = W_1H_1 + W_2H_2, \tag{3}$$

where W_1H_1 is the almost orthogonal part with rank k' and W_2H_2 are e regular NMF components. Following the same idea as in section 2.2, we obtain :

$$H_1 = W_1^T(V - W_2H_2), \quad \text{iff} \quad W_1^TW_1 = I\ . \tag{4}$$

We then propose the *structured projected NMF* cost function:

$$\min_{W_1,W_2,H_2\geq 0} ||V - W_1W_1^T(V - W_2H_2) - W_2H_2||^2. \tag{5}$$

As in [?], e is kept smaller than k'. The goal here is to focus most of the energy in the orthogonal part to benefit from the sparse decomposition property of PNMF.

3. CONSTRUCTION OF THE DICTIONARY

3.1 Database

The dataset is taken from medley-dB [9], it is composed of polyphonic real-world music excerpts. It has 122 music signals and 77 of them contain percussive instruments, harmonic instruments and vocals. The signals that do not contain a percussive part are not part of the evaluation. We will be using the song of the genre, *Singer/Songwriter* (17 songs), *Pop* (10 songs), *Rock* (20 songs), *Jazz* (11 songs), *Electronic/Fusion* (13 songs) and *World/Folk* (6 songs).

3.2 Supervised NMF for source separation

The NMF model is:

$$V \approx \tilde{V} = WH. \tag{6}$$

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. This dictionary can then be used to extract the percussive instruments from a mixture [4]. However, building a dictionary specific to an instrument that performs well on a large database is a complicated problem. Here we build genre specific drum dictionary using the medley-dB database. Using dictionary specific to the genre of music allows us to have smaller dictionaries that a more specific to the signal to decompose. It grants us lower computation time and better separation score The dictionary are build has follow. For every song of the medley-dB database, we perform and NMF with $k = 100$ on the drum signals. The $Wtrain$ matrices are then concatenated depending on the genre of the song.

4. PAGE SIZE

The proceedings will be printed on portrait A4-size paper (21.0cm x 29.7cm). All material on each page should fit within a rectangle of 17.2cm x 25.2cm, centered on the page, beginning 2.0cm from the top of the page and ending with 2.5cm from the bottom. The left and right margins should be 1.9cm. The text should be in two 8.2cm columns with a 0.8cm gutter. All text must be in a two-column format. Text must be fully justified.

5. TYPESET TEXT

5.1 Normal or Body Text

Please use a 10pt (point) Times font. Sans-serif or non-proportional fonts can be used only for special purposes, such as distinguishing source code text.

The first paragraph in each section should not be indented, but all other paragraphs should be.

5.2 Title and Authors

The title is 14pt Times, bold, caps, upper case, centered. Authors' names are omitted when submitting for double-blind reviewing. The following is for making a camera-ready version. Authors' names are centered. The lead author's name is to be listed first (left-most), and the co-authors' names after. If the addresses for all authors are the same, include the address only once, centered. If the authors have different addresses, put the addresses, evenly spaced, under each authors' name.

5.3 First Page Copyright Notice

Please include the copyright notice exactly as it appears here in the lower left-hand corner of the page. It is set in 8pt Times.

5.4 Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

6. FIRST LEVEL HEADINGS

First level headings are in Times 10pt bold, centered with 1 line of space above the section head, and 1/2 space below it. For a section header immediately followed by a subsection header, the space should be merged.

6.1 Second Level Headings

Second level headings are in Times 10pt bold, flush left, with 1 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

6.1.1 Third and Further Level Headings

Third level headings are in Times 10pt italic, flush left, with 1/2 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

Using more than three levels of headings is highly discouraged.

7. FOOTNOTES AND FIGURES

7.1 Footnotes

Indicate footnotes with a number in the text. [1] Use 8pt type for footnotes. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a 0.5pt horizontal rule.

7.2 Figures, Tables and Captions

All artwork must be centered, neat, clean, and legible. All lines should be very dark for purposes of reproduction and art work should not be hand-drawn. The proceedings are not in color, and therefore all figures must make sense in black-and-white form. Figure and table numbers and captions always appear below the figure. Leave 1 line space between the figure or table and the caption. Each figure or table is numbered consecutively. Captions should be Times 10pt. Place tables/figures in text as close to the reference as possible. References to tables and figures should be capitalized, for example: see Figure 1 and Table 1. Figures and tables may extend across both columns to a maximum width of 17.2cm.

String value	Numeric value
Hello ISMIR	2015

Table 1. Table captions should be placed below the table.

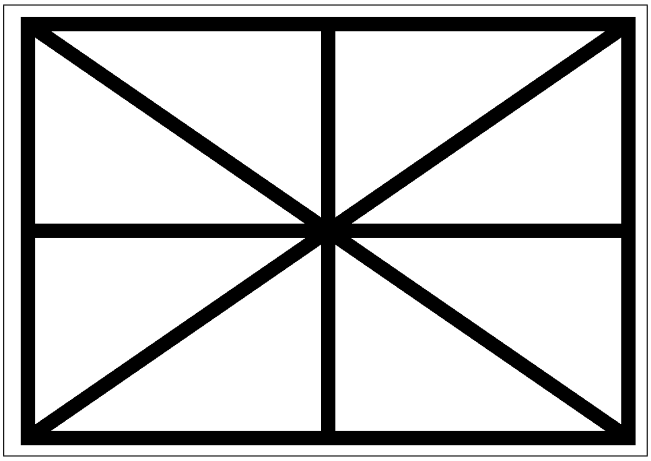

Figure 1. Figure captions should be placed below the figure.

8. EQUATIONS

Equations should be placed on separate lines and numbered. The number should be on the right side, in parentheses, as in Eqn (7).

$$E = mc^2 \tag{7}$$

9. CITATIONS

10. REFERENCES

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

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

[3] Xabier Jaureguiberry, Pierre Leveau, Simon Maller, and Juan José Burred. Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation. In *Proc. of IEEE ICASSP*, pages 5–8, 2011.

[1] This is a footnote.

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

[5] 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.

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

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

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

[9] 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.

[10] 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*, 5(6):1192–1204, 2011.

[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.