DRUM TRANSCRIPTION USING PARTIALLY FIXED NON-NEGATIVE MATRIX FACTORIZATION

Chih-Wei Wu, Alexander Lerch

Georgia Institute of Technology
Center for Music Technology
840 McMillan St. Atlanta GA 30332

ABSTRACT

In this paper, a drum transcription algorithm using partially fixed non-negative matrix factorization is presented. The proposed method allows users to identify percussive events in complex mixtures with a minimal training set. The algorithm decomposes the music signal into two parts: percussive part with pre-defined drum templates and harmonic part with undefined entries. The harmonic part is able to adapt to the music content, allowing the algorithm to work in polyphonic mixtures. Drum event times can be simply picked from the percussive activation matrix with onset detection. The system is efficient and robust even with a minimal training set. The recognition rates for the ENST dataset vary from 56.7 to 78.9% for three percussive instruments extracted from polyphonic music.

Index Terms— NMF, MIR, Drum Transcription, Automatic Music Transcription

1. INTRODUCTION

Automatic music transcription is an intensively researched area in Music Information Retrieval (MIR). The reliable extraction of a score (or a score-related representation) from the audio signal is the core technology of a large number of applications in fields such as music education, systematic musicology, and music visualization. Furthermore, a reliable transcription would enable high-level representations of music signals with the potential of improving virtually any MIR task.

A complete transcription system comprises many related sub-tasks such as multi-pitch detection, onset detection, instrument recognition, and rhythm extraction [1]. While the main focus has been mostly on pitched instruments, a considerable amount of publications deal with the transcription of percussive sounds in mixtures of tonal and percussive instruments. The drum track in popular music conveys information about tempo, rhythm, style, and — at least partly — the structure of a song. A drum transcription system enables applications in active listening [2], music education, and interactive music performance [3].

This study explores the application of the popular transcription method of Non-negative Matrix Factorization (NMF) for drum transcription from polyphonic music. The paper is structured as follows: Section 2 provides an overview of the research in this area. In Section 3 we present our approach; evaluation results are being presented and discussed in Section 4. Section 5 provides a summary, conclusion, and directions of future work.

2. RELATED WORK

Early attempts to transcribe percussive sounds mainly focused on the classification of signals containing solely drum sounds. For these systems, standard approaches with a feature extractor and a subsequent classification engine are able to produce results with high accuracy [4]. For many real-world applications, however, the input file often comprises a mixture of percussive and harmonic sound sources. For most use cases, a drum transcription system is expected to work on this mixture of sounds instead of exclusively on percussive sounds. Gillet and Richard divide systems for the drum transcription from mixtures into three categories [5]: (i) *segment and classify*, (ii) *separate and detect*, and (iii) *match and adapt*.

Systems of the first category (*segment and classify*) usually segment the audio signal into a series of events by applying automatic onset detection and extract various features from time or spectral domain. Each event segment is then classified based on the extracted features. This approach seems to perform well when the features are well chosen [6, 7]. However, a sufficient amount of training data and carefully adjusted pre-processing is required in order to get good results. When working with single-label classifiers, the number of drum classes increases substantially due to the possibility of simultaneous events.

The second type of approaches (*separate and detect*) is based on the assumption that the music signal is a superposition of different sound sources. By decomposing the signal into source templates with corresponding activation functions, the music content could be transcribed by identifying the templates and analysing the activities for each template. Different methods such as Independent Subspace Analysis [8], Prior Subspace Analysis [9], and Non-negative Matrix Factorization (NMF, see below) [10, 11] fall into this category. The advantage of these approaches is that they usually are easier to interpret since most of the decompositions are carried out on the spectrogram of the signal. Furthermore, the handling of simultaneous and overlapping events is inherent to the approach. However, one potential problem in the context of NMF with a pre-determined dictionary matrix is whether or not the templates are representative enough. Another difficulty is the determination of the rank required for the decomposition process.

The third type of approaches (*match and adapt*) uses pre-trained templates to detect drum events [12]. The templates are searched for the closest match and adapted in an iterative process.

3. METHOD

3.1. Algorithm Description

In this paper, we propose a method using partially fixed NMF to transcribe drum events in polyphonic signals. The idea of using NMF with prior knowledge of the target source within the mixture has previously been applied to source separation tasks [13], and multipitch analysis [14]. The method described here is based similar ideas but with different emphasis: (i) we focus on a real world scenario in which users only have limited amount of training samples that are possibly different from the target source, and (ii) we propose to use a small dictionary matrix which is both efficient and easily interpretable.

The basic concept of NMF is to approximate a matrix V with matrices W and H as $V \approx WH$ with non-negativity constraints. Given a $m \times n$ matrix V, NMF will decompose the matrix into the product of a $m \times r$ dictionary matrix W and an $r \times n$ activation matrix H, with r being the rank of the NMF decomposition. In most audio applications, V is the spectrogram to be decomposed, W contains the magnitude spectra of the salient components, and H indicates the activation of these components with respect to time [15]. The matrices W and H are estimated through an iterative process that minimizes a distance measure between the target spectrogram V and its approximation [16].

When NMF is applied to the task of music transcription, typically the following challenges have to be faced: First, the number of sound sources and notes within a music recording is usually unknown. It is therefore difficult to determine a suitable rank r in order to obtain a clear differentiation of the decomposed components in the dictionary matrix. Second, it can be hard to identify the corresponding instrument of every component in the dictionary matrix W. This problem becomes more severe when the rank is selected too high or too low. Third, when multiple similar entries exist in the dictionary matrix, the corresponding activation matrix could be activated at these entries simultaneously, which in turn increases the difficulty of intuitively interpreting the results. Different methods have been proposed in previous studies to address these issues. Helen and Virtanen trained an SVM to separate drum components from the harmonic components; the rank number was derived empirically during the factorization process [17]. The identified drum components and their corresponding activities could later be used to reconstruct the drum signal, resulting in a system for drum source separation. Their approach requires a significant amount of training data for the classifier and, more importantly, the results can be expected to be very sensitive to choice of rank.

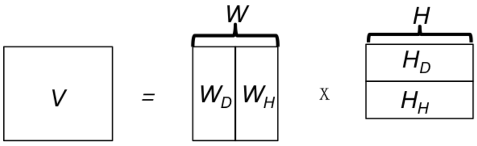

Fig. 1. Illustration of the factorization process. Subscript D: drum components H: harmonic components.

Yoo et al. proposed a co-factorization algorithm [18] to simultaneously factorize a drum track and a polyphonic signal. They used the dictionary matrix from the drum track to identify the drum components in the polyphonic signal. This approach ensures that the drum components in both dictionary matrices are estimated only from the drum track, resulting in proper isolation of the harmonic components from the drum components. Since their system aims at drum separation they can work at very high ranks. For drum transcription, however, the approach is not directly applicable because of the probable lack of interpretability of the dictionary matrix.

Nevertheless, their work inspired our approach to drum transcription. Figure 1 visualizes the basic concept from the work of Yoo et al.: the matrices W and H are split into the matrices W_D and W_H, and H_D and H_H, respectively. Instead of using co-factorization, however, we propose to initialize the matrix W_D with drum templates and to not modify it during the factorization process. Matrices W_H, H_H, and H_D are initialized with random numbers. The distance measure used in this paper is KL-divergence, in which $D_\mathrm{KL}(x \mid y) = x \cdot \log(x/y) + (y - x)$. The cost function as shown in (1) is minimized by applying gradient decent and multiplicative update rules as described in [16], and the matrices W_H, H_H, and H_D will be updated according to (2)–(4).

$$J = D_\mathrm{KL}(V \mid W_\mathrm{D}H_\mathrm{D} + W_\mathrm{H}H_\mathrm{H}) \quad (1)$$

$$H_\mathrm{D} \leftarrow H_\mathrm{D}\frac{W_\mathrm{D}^T(V/(W_\mathrm{D}H_\mathrm{D} + W_\mathrm{H}H_\mathrm{H}))}{W_\mathrm{D}^T} \quad (2)$$

$$W_\mathrm{H} \leftarrow W_\mathrm{H}\frac{(V/(W_\mathrm{D}H_\mathrm{D} + W_\mathrm{H}H_\mathrm{H}))H_\mathrm{H}^T}{H_\mathrm{H}^T} \quad (3)$$

$$H_\mathrm{H} \leftarrow H_\mathrm{H}\frac{W_\mathrm{H}^T(V/(W_\mathrm{D}H_\mathrm{D} + W_\mathrm{H}H_\mathrm{H}))}{W_\mathrm{H}^T} \quad (4)$$

To summarize, the method consists of the following steps:

1. Construct a $m \times r_\mathrm{D}$ dictionary matrix W_D, with r_D being

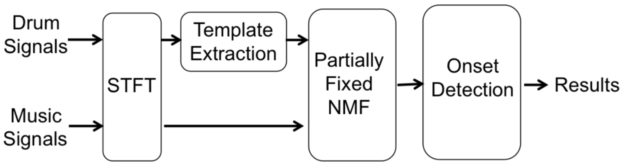

Fig. 2. Flowchart of the drum transcription system

the number of drum components to be detected.

2. Given a pre-defined rank r_{H}, initialize a $m \times r_{\mathrm{H}}$ matrix W_{H}, a $r_{\mathrm{D}} \times n$ matrix H_{D} and a $r_{\mathrm{H}} \times n$ matrix H_{H}.
3. Normalize W_{D} and W_{H}.
4. Update H_{D}, W_{H}, and H_{H} using (2)–(4).
5. Calculate the cost of the current iteration using (1).
6. Repeat step 3 to step 5 until convergence.

The time positions of the drum events can then be extracted by applying a simple onset detection on the rows of matrix H_{D}.

3.2. Implementation

Figure 2 shows the flow chart of the implemented system. Since the NMF is based on a constructive assumption of multiple sources, the representation of the signal must be superimposable. Therefore, a magnitude spectrogram is used to instead of other audio specific features. The STFT of the signals will be calculated using a window size and a hop size of 2048 and 512 with a sampling frequency of 44.1 kHz. A pre-trained dictionary matrix will be constructed from the training set, consisting of isolated drum sounds. Next, the partially fixed NMF will be performed with rank $r = r_{\mathrm{D}} + r_{\mathrm{H}}$ as described above. Finally, the activation Matrix H_{D} is evaluated to determine the onset positions and their corresponding classes.

As mentioned above, the dictionary matrix W_{D} is created by extracting a template spectrum from isolated training drum samples. The template magnitude spectrum is a median spectrum of all individual events of one drum class in the training set. The length of each event is approximately 80 ms. The templates are extracted for the three classes Hi-Hat (HH), Bass Drum (BD) and Snare Drum (SD).

High values in the activation matrix H_{D} indicate the presence of a drum event. More specifically, the activity difference of each row of the activation matrix could be considered as the onset novelty function of each individual drum. We use a median filter to create an adaptive threshold for peak picking. The implementation of the median filter is shown in (5). The G is the time-varying threshold. Q is a function that extracts the median from previous input signal within a fixed window size. λ is an offset coefficient to control the sensitivity of the threshold. For every track, we set the window size to be 0.1 s and the λ to be 0.12 of the maximum value, respectively.

$$G(t) = \lambda + Q(t) \tag{5}$$

	Dr1	Dr2	Dr3	Total
HH	1942	2145	1813	5900
BD	2140	1488	1378	5006
SD	2165	2079	1994	6238
Total	6247	5712	5185	17144

Table 1. Onset counts in selected data set

4. EVALUATION

4.1. Dataset Description

The experiments have been conducted on the *minus one* subset from the ENST public drum data set [19]. This data set consists of recordings from three different drummers performing on their own drum kits. The set for each drummer contains individual hits, short phrases of drum beats, drum solos, and short excerpts played with the accompaniments. The minus one subset has 64 tracks of polyphonic music, and the sampling rate of every track is 44.1 kHz. Each track in this subset has a length of approximately 70 s with varying style. More specifically, the subset contains various drum playing techniques such as ghost notes, flam, and drag; these techniques are considered difficult to identify with existing drum transcription systems. Since we are only interested in the three classes HH, BD, and SD, tracks missing one of these instruments or featuring specific playing techniques have been discarded, leaving a subset of 53 out of 64 tracks.

The accompaniments are mixed with the drum tracks in the data set without any modification (e.g., no level adjustment). The distribution of onset counts per class per drummer is shown in Table 1. The drum templates have been generated from a different part of the dataset which only contains single hits performed by the same group of drummers. Each track contains 5 to 6 single hits on different drums for each drummer. The onset position of these single hits was determined using the annotated ground truth.

4.2. Evaluation Procedure

We evaluate two different combinations of training and test data: First, we use training samples from all three drummers to train the drum dictionary matrix, and test the system on all 53 tracks; second, we investigate the cross-performer accuracy. In the latter scenario, the training samples are selected only from one drummer, and tested on the other drummers' recordings. This scenario should be similar to a real-world use case for which the trained drum sounds are not necessarily similar to the drum sounds in the target signals.

The evaluation metrics follow the standard calculation of the precision (P), recall (R), and F-measure (F). An onset is considered to be a match with the ground truth if the time deviation between the annotated and detected onset is less or equal to 50 ms.

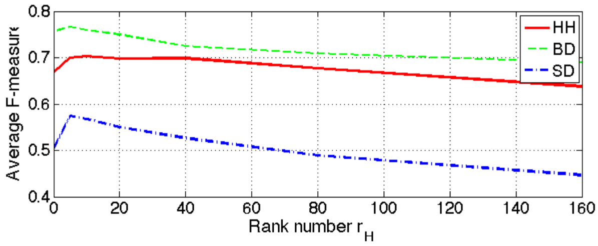

Fig. 3. Average F-measure versus harmonic rank r_{H}

4.3. Evaluation Results

In an initial test to determine the rank r_{H} of the algorithm, $r_{\mathrm{H}} = 5, 10, 20, 40, 80, 160$ have been tested. The resulting individual F-measures are shown in Figure 3. A general trend of decreasing performance with increasing r_{H} can be observed, especially for lower frequency sounds such as SD and BD. Based on this observation, a rank number $r_{\mathrm{H}} = 10$ is chosen in current setup.

Instrument	HH	BD	SD
P	0.681	0.755	0.634
R	0.727	0.827	0.513
F	**0.703**	**0.789**	**0.567**

Table 2. Transcription results using all training templates

Table 2 shows the results when the system was trained with the recordings of all three drummers. Gillet and Richard reported an F-measure of 77.7%, 65.0% and 64.8% for HH, BD, and SD for the same dataset, using a sophisticated approach requiring a significant amount of training data [5]. We observe that our systems performs better performance on BD, but slightly worse for SD and HH.

In order to investigate the dependency of our approach with respect to the similarity of training and test drum sounds, we conduct a cross-performer evaluation as mentioned in Section 4.2. The results, listed in Table 3, show a simple trend: the test set containing drummer 2 nearly always gives the best results, regardless of the training set. Also, when training with different drummer's recordings, the F-measure from HH and SD are mostly within the same range as the results reported in Table 2 except for BD. This could be due to the fact that Bass Drum of drummer 2 is easy to detect. These results indicate that the presented algorithm is relatively robust against differences between the drum template and the sound of the drum to be detected. This would allow to construct a template from different sound sources independent of the recording to be analyzed allowing more general applications. However, the performance of this setup still needs to be confirmed with a cross-data set validation.

Training		Dr1	Dr2	Dr3	Avg.
Testing		Dr2+Dr3	Dr1+Dr3	Dr1+Dr2	
HH	P	0.661	0.662	0.660	0.661
	R	0.747	0.719	0.726	0.731
	F	**0.701**	0.689	0.691	0.694
BD	P	0.811	0.641	0.859	0.770
	R	0.905	0.742	0.915	0.854
	F	0.855	0.687	**0.886**	0.810
SD	P	0.724	0.543	0.727	0.665
	R	0.538	0.429	0.539	0.502
	F	0.617	0.479	**0.619**	0.572

Table 3. Transcription results of cross-performer validation.

5. CONCLUSION

We have presented a drum transcription system for polyphonic music using partially fixed NMF. This method uses a partially pre-trained dictionary matrix to decompose the target signal and to estimate the activation matrix. The evaluation results show that this method is able to achieve 56.7 to 78.9% F-measure for detecting 3 classes from complex mixtures of music.

The presented method has the following advantages: First, the fixed dictionary matrix in the model makes it easier to interpret the corresponding activation matrix for transcription tasks. Second, simultaneous sounds can be detected separately without the need of training extra classes. Third, adjustment of the parameter r_{H} allows the algorithm to adapt to different different types of polyphonic music. Fourth, cross-performer evaluation results indicate a robustness against template mismatches, possibly allowing the application in situations with minimum prior knowledge. Last but not least, the approach only requires a few drum samples to train the dictionary matrix, and the evaluation results indicate that the performance is comparable with state-of-the art methods at lower algorithmic complexity.

Possible directions for future work are: a comparison between this approach and Probabilistic Latent Component Analysis (PLCA) [20]. We will also investigate means to iteratively adapt the template during the decomposition as a way of improving the current method. Furthermore, the automatic estimation of r_{H} for any given signal using a probabilistic approach similar to [21] might be a solution to rank selection. Finally, different penalty terms for the cost function, such as sparsity, temporal continuity [22], or rank r_{H} might be taken into account for better adjustment of the current method. To reach the goal of a complete drum transcription system for polyphonic music, however, more factors such as playing techniques and more drum classes still need to be addressed in the future.

REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri, "Automatic

music transcription: Challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407—434, Dec. 2013.

[2] Kazuyoshi Yoshii, Masataka Goto, and Kazunori Komatani, "Drumix: An audio player with real-time drum-part rearrangement functions for active music listening," *IPSJ Digital Courier*, vol. 3, pp. 134—144, 2007.

[3] Gil Weinberg, Aparna Raman, and Trishul Mallikarjuna, "Interactive jamming with shimon: a social robotic musician," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 2009, pp. 233—234, ACM.

[4] Perfecto Herrera, Amaury Dehamel, and Fabien Gouyon, "Automatic labeling of unpitched percussion sounds," in *Proceedings of the 114th Audio Engineering Society Convention*. Mar. 2003, AES.

[5] Olivier Gillet and Gaël Richard, "Transcription and separation of drum signals from polyphonic music," *Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529—540, Mar. 2008.

[6] Olivier Gillet and Gaël Richard, "Automatic transcription of drum loops," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004, pp. iv–269–iv–272 vol.4, 00062.

[7] Christian Dittmar, "Drum detection from polyphonic audio via detailed analysis of the time frequency domain," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2005.

[8] Derry FitzGerald and Bob Lawlor, "Sub-band independent subspace analysis for drum transcription," in *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, Hamburg, 2002.

[9] Derry FitzGerald, Bob Lawlor, and Eugene Coyle, "Drum transcription in the presence of pitched instruments using prior subspace analysis," in *Proceedings of the Irish Signals & Systems Conference (ISSC)*, Limerick, 2003.

[10] Arnaud Moreau and Arthur Flexer, "Drum transcription in polyphonic music using non-negative matrix factorisation," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2007, pp. 353—354.

[11] David S Alves, Jouni Paulus, and José Fonseca, "Drum transcription from multichannel recordings with non-negative matrix factorization," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Glasgow, 2009.

[12] Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G Okuno, "Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression," *Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 333—345, Jan. 2007.

[13] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of the 7th international conference on Independent component analysis and signal separation*, 2007, pp. 414–421.

[14] Stanislaw A. Raczyski, Nobutaka Ono, and Shigeki Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, 2007.

[15] Paris Smaragdis and Judith C Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, 2003, IEEE.

[16] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, 2000.

[17] Marko Helen and Tuomas Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Talya, 2005.

[18] Jiho Yoo, Minje Kim, Kyeongok Kang, and Seungjin Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, 2010, pp. 1942—1945, IEEE.

[19] Olivier Gillet and Gaël Richard, "Enst-drums: an extensive audio-visual database for drum signals processing," in *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, 2006.

[20] Paris Smaragdis, Cedric Fevotte, Gautham J. Mysore, Nasser Mohammadiha, and Matthew Hoffman, "Static and Dynamic Source Separation Using Nonnegative Factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, May 2014.

[21] Mikkel N. Schmidt and Morten Mø rup, "Infinite nonnegative matrix factorization," in *European Signal Processing Conference (EUSIPCO)*, 2010.

[22] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.