NON-NEGATIVE MATRIX FACTORIZATION FOR DRUM SOURCE SEPARATION AND TRANSCRIPTION

Jon Downing
University of Rochester
jdowning@ur.rochester.edu

ABSTRACT

This paper presents a system for transcription and source separation of polyphonic drum recordings. Such a system may find applications in music education, music production, or entertainment. The system's methods for detection and decomposition are based on the well-known Non-Negative Matrix Factorization (NMF) approach. The basic multiplicative update rules are modified to capture the spectral variation over time of the percussive sounds per frame by using semi-adaptive update rules for the spectral templates. Additionally, two dictionary atoms are stored for each drum sound contained in the mixture, corresponding to the initial transient and steady-state decay of the drum sound. State-of-the-art onset detection methods are examined and applied to the initial decomposition. The proposed modification is shown to improve the f-score of the transcription given an identical onset detection function. We compare the transcription statistics over a dataset generated from acoustic and electronic drum samples.

1. INTRODUCTION

With the advent of digital recording technology, new possibilities in music performance, production, and transcription have presented themselves. The work in this paper is centered around the transcription and separation of single drum instruments when presented with a monaural, polyphonic recording of a drum kit performance. Thus, we are concerned with the topics of both automatic music transcription and musical source separation. Each of these is a major topic of research in the field of Music Information Retrieval (MIR) [1, 2].

The goals and practical applications of this work are well summarized in [2] and reiterated below. In a typical recording situation, each drum is recorded with a dedicated microphone to allow for separate treatment and processing of these individual drum signals; nonetheless, cross-talk between microphones is often unavoidable, and in a studio with a limited budget, a dedicated microphone may not be available for each drum. Thus, there are considerable advantages to a single-microphone solution in which single drums are isolated and transcribed in real-time from a monaural source recording, and the proposed method may find use in a piece of a music production software for this purpose. Additionally, a number of commercially available educational video games, such as RockSmith and BandFuse, are available, allowing users to practice drums as their performance and timing is assessed in real-time. However, all such systems are dependent on MIDI-triggered drum sounds, and employ an electronic drum kit. The method described in this paper may enable the development of a drum training system which allows the user to perform on a real, acoustic drum kit, while providing a similar level of feedback and performance assessment. Finally, the system may be applied to more traditional transcription purposes in the event that a source recording of the isolated drum kit is available; for example, to aid in transcription of a drum solo, or of the entire drumbeat of a song if the isolated drum recording is available.

As in [2], the proposed method is dependent upon a training phase in which each of the drum sounds contained in the transcription mixture is recorded and analyzed in isolation, to establish spectral templates to be matched in the transcription phase. This requirement is easily satisfied in each of the three application scenarios mentioned above by performing a ”sound check” in which each individual drum is struck a number of times in isolation. While there are a multitude of different types of drums, and an infinite number of drum kit configurations, there are three percussive instruments in particular which are found in virtually all drum kits and make up the basis of the drum performances in much of contemporary popular music: the kick drum, snare drum, and hi-hat. For this reason, we focus on the transcription and decomposition of these three drum sounds in this paper.

The kick drum is a large drum, typically struck with a foot pedal, and produces a low-pitched resonant sound. The snare drum is smaller, and features a set of metal wires stretched over the bottom drum head, lending it a brighter sound with a higher-pitched resonance than the kick drum. The hi hat is a pair of cymbals struck of which the top cymbal is struck during performance. A foot pedal enables switching between an open hi hat configuration, in which the cymbals ring out against each other, and closed configuration, in which the cymbals are held tightly together. The hi hat produces a click-like sound with a rapid decay in closed mode, and a bright, splashy sound with a long decay in open mode. While a naive approach to transcription of these sounds might simply classify drum occurrences by their spectral centroid, this fails in practice because drum sounds often coincide with each other in a polyphonic drum performance. For this reason, the proposed method first performs source separation on the monaural mixture of drum sounds, then transcribes each drum's performed rhythm in isolation.

The remainder of this paper adheres to the following structure. In Sec. 2, the current state-of-the-art in source separation for drum transcription using NMF is summarized. In Sec. 3, the novel method is proposed, and its relation to prior work is examined. In Sec. 4, experiments are performed on the proposed method, and results of the novel approach are compared to those of other state-of-the-art methods. Finally, Sec. 5 summarizes these results and the significance of the findings.

2. PRIOR WORK

This section summarizes some state-of-the-art techniques for drum source transcription and separation. Since the decomposition and transcription processes are performed separately in our approach, we examine methodologies for each separately. We discuss prior work in drum source separation, focusing on NMF-based decompositions, since these are most closely related to the proposed method.

NMF is a learning algorithm for the decomposition of a nonnegative matrix X. Lee and Seung provide a discussion of the general form of the NMF problem along with proofs of convergence for proposed solutions to the problem in [3]. In general, given a nonnegative matrix X, NMF attempts to find non-negative matrices B and H such that

$$X \approx BH \tag{1}$$

Given a matrix X with dimensions n x m, then, we seek a factorization consisting of a n x r matrix W and an r x m matrix H, with W and H both nonnegative. We typically choose a value of r much smaller than either n or m, such that B and H yield a compressed representation of the initial data. We see that each column of X can be approximated by $x \approx B_h$, where x and h are corresponding columns of X and H. Thus we can view each column of X as a linear combination of the columns of B weighted by components of h. Inspired by this interpretation, the columns of B are typically referred to as "basis vectors", while the rows of H are often called "activations." It is important to note that a good approximation of the original data can be obtained only if the basis vectors uncover latent structure in the data, since we typically have far more data vectors in X than we have basis vectors in B.

Typically, the initial matrix X is decomposed by minimizing a divergence function between X and the reconstructed matrix BH. A common choice for the cost function is the Kullback-Leibler (KL) divergence, which is given for matrices A and B as

$$D(A||B) = \sum_{ij}(A_{ij} - B_{ij})^2 \tag{2}$$

The KL divergence can be minimized by performing iterative multiplicative updates on B and H according to the following equations as shown in [3]:

$$B \leftarrow B \cdot \frac{\frac{X}{BH}B^T}{1B^T} \tag{3}$$

$$H \leftarrow H \cdot \frac{B^T\frac{X}{BH}}{B^T 1} \tag{4}$$

These rules additionally enforce the non-negative constraints imposed on B and H.

In the application of NMF to the decomposition of musical spectra, the initial data matrix X is typically a magnitude spectrogram of the musical performance obtained via STFT. The matrix B, then, stores spectral templates of musical events in its columns, while the rows of H contain temporal activation weightings for the corresponding spectral templates. Individual musical voices can then be isolated from the spectrogram mixture by multiplying single columns of B with corresponding rows of H, yielding a new spectrogram of the isolated source.

A modification of the Lee and Seung update rules is presented in [2], which makes use of semi-adaptive bases to apply NMF to drum source separation. In this method, he matrix B is initialized with a learned spectral basis matrix B_p. Each basis vector in B_p is obtained by taking the STFT of the training recording for each drum source, and simply averaging across the time dimension. In the semi-adaptive approach, rather than leaving B_p fixed throughout the decomposition, or allowing it to freely iterate towards B according to the standard update rules, a blending parameter, α, is introduced to provide a mixture of these two approaches. Eqn (5) shows the update step which replaces Eqn (3), and illustrates how the spectral content of B is weighted towards its initial value, B_p with high α, and reverts to the NMF decomposition B with low α. Eqn (6) shows the calculation of α, which decreases as the current NMF iteration, k, approaches the iteration limit, K, based on a user-selected parameter, β.

$$B = \alpha \cdot B_p + (1 - \alpha) \cdot B \tag{5}$$

$$\alpha = (1 - \frac{k}{K})^\beta \tag{6}$$

In order to best capture the rapidly changing spectral-temporal characteristics of percussive sound sources, the methodology of [2] calls for the application of the NMF algorithm to each individual spectral frame of the monaural drum recording, with the vectors B and H reverting to their initial values following NMF decomposition of the previous frame.

3. PROPOSED METHOD

We propose a novel method for establishing the basis matrix, B_p. The rank of the NMF decomposition is expanded from 3 to 6: we introduce separate templates for drum "heads" and "tails" as in [5]. We first run onset detection on the training data for each drum. The head template, B_H, is taken as the time average across spectrogram frames containing onsets, while the tail spectrum for each drum is simply the time-average across the remaining spectrogram frames. Now semi-adaptive NMF is applied on a

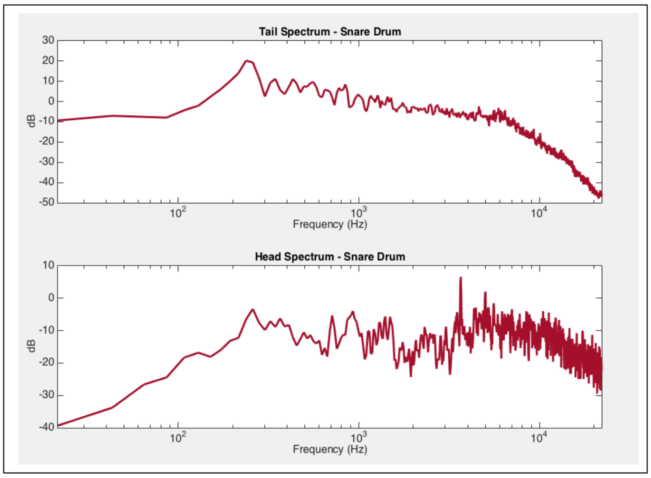

Figure 1. Comparison of head and tail spectra for snare drum.

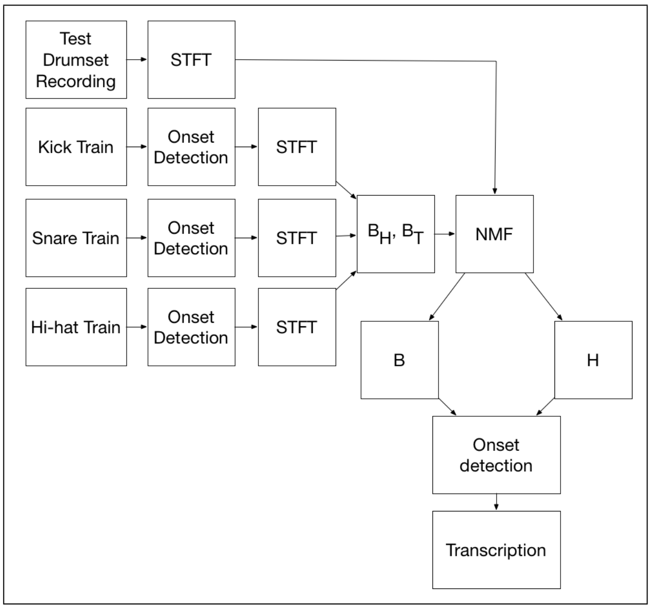

Figure 2. Block diagram of proposed method.

per-frame basis to the magnitude spectrogram of the test data, in order to perform the source separation as in [2]. Onset detection is then applied to reconstructed spectrograms of each drum, taken as the sum of the head and tail reconstructions.

Many mistaken onsets in NMF-based drum transcription are caused by cross talk between the activations [2, 4, 5]. This is because drum attack transients typically manifest as short bursts of broadband noise; thus, if no spectral template is a very close match for the transient, other templates will be activated to attempt a better approximation. In order to obtain the most salient activation information in H, we should begin with spectral templates for drum attacks which are unique to the corresponding drum.

We conjecture that expanding the rank of the NMF by introducing the head templates will reduce the number of mistaken onsets due to crosstalk. Since we begin each NMF decomposition with templates explaining the attack of each drum hit, we will be less likely to see these transient weightings assigned to the activations of templates corresponding to the incorrect drum. In a sense, the head template should ”absorb” much of the activation energy which would otherwise be distributed among all rows of H in an attempt to approximate the broadband burst of noise corresponding to the drum attack.

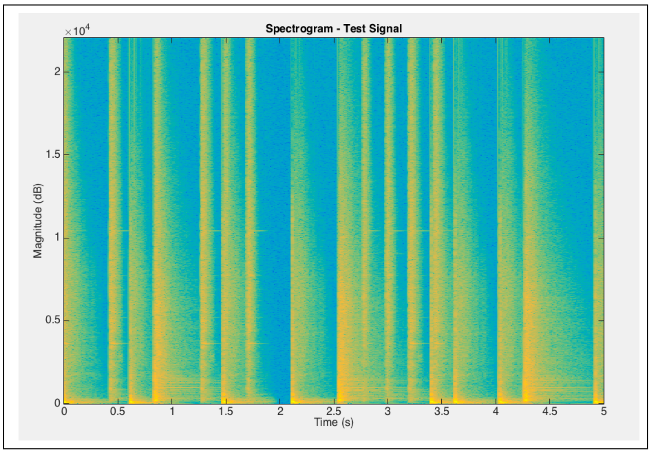

Figure 3. Spectrogram of a test mixture containing kick, snare, and hi-hat.

4. EXPERIMENTS

Experiments were conducted comparing the rank 3 decomposition presented in [2] with the novel approach employing head and tail templates for each drum. We examine the rank 3 decomposition with fixed templates, semi-adaptive templates, semi-adaptive heads with fixed tails, and semi-adaptive tails with fixed heads. We also include blind NMF results for comparison. The data set for these results has its origin in [6]. The expanded data-set, containing onset annotations and synthetic drum patterns in addition to the original acoustic recordings, was compiled in [2] and it is this set on which we present our results. There are a total of 10 minutes of audio, containing 33 separate drum sequences with 3741 annotated onsets. Following the findings of [2], we selected parameters of $H = 512$ samples hop-size, $N = 2048$ bins spectrum size, $K = 25$ NMF iterations, and $\beta = 4$. For the onset detection function, we apply a simplified version of [1], to the reconstructed magnitude spectrogram for each drum, based solely on the first-order difference of the half-wave rectified amplitude and simple thresholding. We used a threshold of 0.4 for the normalized first-order difference to determine an onset. From the onset detection, we calculate the location of the onsets in units of seconds, and compare the results for each tail component to the

Experimental results are presented in Table 1. We see that the blind NMF approach compares poorly to the source-informed methods. With semi-adaptive rank 3 implementation, we acheive and f-score of 0.90. Although these results do not equal the f-score of 0.95 obtained using this method in [2], this is probably due to the less sophisticated onset detection method employed in our experiments. If we expand the rank to 6 and obtain separate head and tail

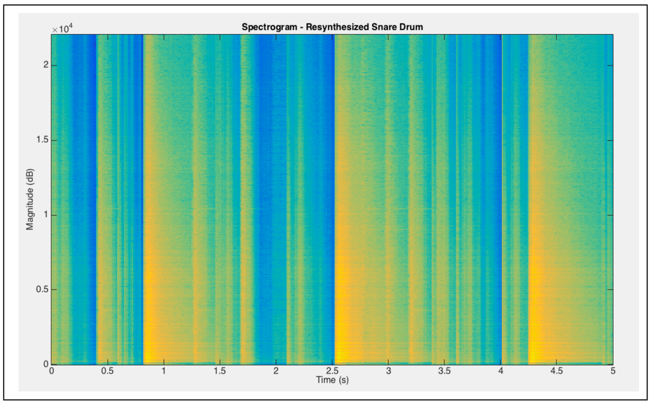

Figure 4. Spectrogram of the separated snare drum for the test mixture from the previous figure.

templates, as proposed in our paper, both fixed and semi-adaptive NMF yield a significantly lower f-score than the semi-adaptive rank 3 method. However, with fixed head templates, we achieve results essentially equivalent to the semi-adaptive rank 3 method.

We can conclude that simply expanding the rank of the spectral templates in semi-adaptive NMF for drum transcription to contain head and tail templates actually results in poorer transcription performance. It is believed that the semi-adaptive head templates will explain onsets for each of the drums fairly well by the end of the iteration limit, and that any broadband transient in the spectrogram will end up being distributed among each of these adaptive templates. These head templates effectively consume spectral information relating to noisy parts of the signal and distribute it equally among the drum components; this explains why this approach yields a low precision, because many of the peaks in activation are distributed too sparsely across the components to trigger the correct onset. It has been observed that the semi-adaptive learned spectra for the head templates look very similar for frames containing coinciding onsets of multiple drums, supporting this conclusion. However we also note that by fixing the head templates and allowing the tails to adapt, we preserve the most distinguishing information about the attack of each drum, throughout the iteration cycle. The result is the that there is less crosstalk overall than semi-adaptive heads, and the results approach the state of the art results of [2]. It is interesting to note that allowing the tail spectra to adapt, however, still results in a measurable improvement over fixed NMF. Further work is needed to examine if there are any unobserved benefits to be gained from this expanded rank NMF decomposition with fixed heads. We are particularly interested to compare the perceptual quality of the audio source separation for each of the examined methods.

5. CONCLUSION

A novel method for polyphonic drum source separation and transcription was proposed. The method uses two spectral templates per drum source to model the attack and decay of the percussive sounds separately. The novel approach was tested on a standardized data set. The f-score were achieved with the modifications proposed in this paper equal those obtained using standard semi-adaptive NMF, provided the head templates of the drums are fixed. This suggests that the rank 6 decomposition approach may be applied with success to polyphonic drum transcription.

Method	Precision	Recall	f-score
Blind NMF	0.76	0.72	0.73
Semi-adaptive (r=3)	0.95	0.88	0.90
Fixed Head/Tail Templates	0.93	0.84	0.87
Semi-adaptive Heads/Tails	0.83	0.87	0.83
Semi-adaptive, Fixed Tails	0.89	0.66	0.72
Semi-adaptive,Fixed Heads	0.95	0.88	0.90

Table 1. Results of experiments.

6. REFERENCES

[1] W. Wang, Y. Luo, J.A. Chambers and S. Sanei. ”Non-Negative Matrix Factorization for Note Onset Detection of Audio Signals,” *Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pp. 447–452, 2006.

[2] C. Dittmar and D. Gartner. “Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition,” *Proceedings of the 17th International Conference on Digital Audio Effects*, 2014.

[3] D. Lee and H. Seung. ”Algorithms for Non-negative Matrix Factorization,” *Advances in neural information processing systems*, Vol. 13, 2001.

[4] S. Tjoa and K. J. Liu. ”Multiplicative Update Rules for Nonnegative Matrix Factorization with Co-occurrence Constraints.” *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, TX, March 2010.

[5] T. Gifford and A. Brown. ”Listening for noise: An approach to percussive onset detection,” *Sound : Space - The Australasian Computer Music Conference*, Sydney, 2008.

[6] E. Battenberg, V. Huang, and D. Wessel, ”Live drum separation using probabilistic spectral clustering based on the itakura-saito divergance,” *Proceedings of the AES 45th Conference on Time-Frequency Processing in Audio*, Helsinki, Finland, 2012.