ABSTRACT

This poster 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. The proposed method is shown to improve the f-score of the transcription given an identical onset detection function by comparing the transcription statistics over a dataset generated from acoustic and electronic drum samples.

BACKGROUND

This work in this paper is centered around the transcription and separation of single drum instruments from single monaural, polyphonic recordings of drum kit performances. In an educational context, this can enable the user to obtain feedback on a drum performance. Additionally, a live drum performance could be transcribed for later analysis or performance. In each of these applications, we require separate subsystems for source separation and transcription; hence, we examine Non-Negative Matrix Factorization (NMF) and onset detection for each task separately.

Non-Negative Matrix Factorization

A common approach to audio source separation is Non-Negative Matrix Factorization, as proposed in [1]. In this approach, the magnitude spectrogram of the signal is decomposed into a lower-rank approximation, consisting of a matrix $B$ of spectral vectors, or templates, and a matrix $H$ of time-varying gain vectors, or activations, for each source. A common approach is to choose a rank for the decomposition and initialize $B$ and $H$ randomly. Then the update rules below, from [1], are applied iteratively until convergence.

$$X \approx BH$$

$$B \leftarrow B \cdot \frac{XH^T B}{BH^T B}$$

$$H \leftarrow H \cdot \frac{B^TXH}{BH^T}$$

Once the NMF decomposition is completed, the outputs $B$ and $H$ can be used to perform onset detection, as proposed in [3].

$$X \approx BH$$

$$B \leftarrow B \cdot \frac{XH^T B}{BH^T B}$$

$$H \leftarrow H \cdot \frac{B^TXH}{BH^T}$$

Onset Detection

Usually, onset detection is performed on the reconstructed spectrograms of the components for transcription. Below is a simple method of extracting onsets from the magnitude spectrogram, employing a 1st-order difference function and simple thresholding.

$$X \approx BH$$

$$B \leftarrow B \cdot \frac{XH^T B}{BH^T B}$$

$$H \leftarrow H \cdot \frac{B^TXH}{BH^T}$$

CONCLUSIONS

Our method makes use of semi-adaptive NMF with spectral templates learned from the isolated drums samples. However, we expand the rank to six templates, consisting of “heads” and “tails,” as proposed in [3]. We obtain the “head” templates by using the onset detection on the training data to determine spectrogram frames corresponding to onsets, and generate our onset templates from only these frames. When semi-adaptive NMF is applied to the test data, it is hoped that crosstalk will be reduced since we now have more salient spectral information about the onsets of each individual drum, which is typically where most crosstalk occurs. The final transcription is done via onset detection on the reconstructed spectrogram of the $B$ and $H$ matrices from the NMF decomposition.

RESULTS

Precision and recall statistics are presented below for the system with no supervision, with semi-adaptive templates from the training data, and with separate head and tail templates from the training data. It was found that the best subset of onset activations to use for transcription were kick snare, snare tail, and hi-hat head. The spectrogram shows the snare drum’s onset prominently in the synthesized snare drum spectrogram. Precision is generally below recall due to mistaken hi hats.

Table 1 – Experimental Results

<table>
<thead>
<tr>
<th>Condition</th>
<th>Precision</th>
<th>Recall</th>
<th>F Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind NMF</td>
<td>0.66</td>
<td>0.69</td>
<td>0.67</td>
</tr>
<tr>
<td>Semi-Adaptive Templates</td>
<td>0.72</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>Fixed Templates Heads and Tails</td>
<td>0.74</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>Semi-Adaptive Heads and Tails</td>
<td>0.74</td>
<td>0.81</td>
<td>0.77</td>
</tr>
</tbody>
</table>

REFERENCES


ACKNOWLEDGEMENTS

Dr. Zhiyao Duan, Dr. Jakob Abeßer, and the University of Rochester Computer Audition class of Fall 2015.