DRUM TRANSCRIPTION IN POLYPHONIC MUSIC USING NON-NEGATIVE MATRIX FACTORISATION

Arnaud Moreau
The Austrian Research Institute
for Artificial Intelligence
Freyung 6/6, A-1010 Vienna, Austria
a.moreau@gmx.net

Arthur Flexer
Institute of Medical Cybernetics
and Artificial Intelligence
Center for Brain Research
Medical University of Vienna, Austria
arthur.flexer@meduniwien.ac.at

ABSTRACT

We present a system that is based on the non-negative matrix factorisation (NMF) algorithm and is able to transcribe drum onset events in polyphonic music. The magnitude spectrogram representation of the input music is divided by the NMF algorithm into source spectra and corresponding time-varying gains. Each of these source components is classified as a drum instrument or non-drum sound and a peak-picking algorithm determines the onset times.

1 INTRODUCTION

The transcription of percussive instruments in music signals is an important step to analyzing its rhythmical content, which is useful in genre classification or beat/meter detection. The detection of drum occurrences is a less difficult task than the transcription of harmonic instruments because percussive instruments in general stay constant in pitch throughout the recording. In the case of pure percussive music different approaches give reasonably accurate results [4], but if pitched instruments are also present in the signal, the task becomes very difficult because they disturb the transcription process. There are two different approaches to the problem that appear in literature: (i) Onset detection based systems [5] first search the input signal for potential drum onsets and then classify them. (ii) Separation based systems [6, 3] first use source separation methods like Independent Subspace Analysis, Prior Subspace Analysis or NMF to divide the input audio signal into source signals and then use some sort of peak-picking algorithm to find relevant onsets. Our system can be seen as an extension of the work presented by Helen and Virtanen [3] who classified the source signals obtained by the NMF algorithm into drum and non-drum signals in order to extract the drum-only signal from the mixture via resynthesis. We, however, not only try to detect the drum sources but also classify them into single drum instruments.

2 METHOD

The magnitude spectrogram $\mathbf{X}$ is computed using a hanning window of size 4096 samples (93 ms) and 75 % overlap. It is therefore a matrix of size $f \times t$ (resolution in the frequency domain × number of frames). One short-time spectrum vector at frame t is modelled as a sum of c components, each having a constant spectrum $\mathbf{S}$ and time-varying gain $A(t)$. This can be written as $\mathbf{X} \approx \mathbf{SA}$. $\mathbf{S}$ is a matrix of size $f \times c$ and $\mathbf{A}$ is a matrix of size $c \times t$. The components are estimated using the NMF algorithm in [3].

The spectra and the gains obtained in the NMF decomposition are used as input to the feature extraction process. All the features we considered are listed in Table 1. The spectral features are computed from the source spectra $S(f)$ and the temporal features are computed from the time-varying gains $A(t)$. Most of those features are commonly used in pattern recognition (for details see [5]). Noise-likeness and percussiveness [6] measure the roughness of the spectrum and the sharpness of attacks in the gain, respectively. Peak time and peak fluctuation are the median and the interquartile range of the durations of the peaks ($A(t) \geq 0.2 \max A(t)$) in the gain. Periodicity [2] measures the correlation of the time-shifted signal. In order to preserve the temporal information, not present in the 10 MFCCs calculated on the source spectrum (which corresponds to 1 frame), we add dynamic MFCCs and ΔMFCCs which are calculated from the magnitude spectrogram ($\mathbf{S} \cdot \mathbf{A}_{peak}$) of the most prominent peak in the time-varying gain. Their means and standard deviations are used as features.

For classification we use a simple one-nearest-neighbor algorithm that works with the scale-invariant mahalanobis-distance. To train our classifier we use 22 polyphonic music excerpts from recordings of different music styles of variable length (5 or 10 seconds) that are divided into 15–25 components by the NMF algorithm. These music ex-

The authors would like to thank Helmut Schönleitner of the cultural center AKKU (http://www.akku-steyr.at) who provided the multichannel recordings that have been used to evaluate our algorithm. The Austrian Research Institute for Artificial Intelligence acknowledges support from the ministries BMUKK and BMVIT.

spectral features	temporal features
spectral centroid	temporal centroid
spectral kurtosis	temporal kurtosis
spectral skewness	temporal skewness
spectral rolloff	crest factor
spectral flatness	peak time
spectral contrast	peak fluctuation
noise likeness	percussiveness
standard deviation	periodicity
10 MFCCs	
20 dynamic MFCCs (mean+std)	
20 dynamic ΔMFCCs (mean+std)	

Table 1. Overview of all features considered in the classification process.

cerpts have no overlap with the test data used for evaluation. All of these components are hand-labelled by listening to them after re-synthesis. The feature extraction is carried out on those components, resulting in 415 feature vectors, distributed as follows: 32 bass drum (BD), 56 snare drum (SD), 34 hihat (HH), 293 non drum (ND). The distribution of the different classes is approximately the same as in the test data. Initial experiments showed that this method outperforms the training based on recordings of isolated drum sounds.

All time-varying gains of drum instruments are fed into the peak picking algorithm. We used a slightly modified version of the one presented in [1].

3 EVALUATION AND DISCUSSION

The proposed system has been evaluated on a song of 1 minute length, which has been divided into 5 second excerpts. The data set contains a total of 260 onsets, which are distributed as follows: 89 BD, 55 SD, 116 HH. It is a recording of contemporary jazz music played by drums, keyboards and a bass guitar. The NMF algorithm was carried out using 15 components ($c = 15$). The transcribed onsets o_t have been compared to the reference onsets o_r using the procedure proposed in [4]. *Precision rate* $R_p = (T - fp)/T$, *recall rate* $R_r = (R - tn)/R$ and *instrument hit rate* $R_h = 1-(fp+tn)/R$ where $fp\ldots$ false positives, $tn\ldots$ true negatives, $T\ldots$ transcribed events and $R\ldots$ reference events, are computed. Results are given in Table 2. The obtained results only serve as illustration of the system's capabilities since they have been computed using only one reference song. Whereas BD events are transcribed very satisfactory, the recognition of SD and HH events is less than optimal. Paulus and Virtanen [4] achieve an average R_h of 96% with their method transcribing drum-only music, so only our result on the BD transcription is acceptable considering the difference in difficulty.

	BD	SD	HH	mean
$R_p\%$	89.47	36.71	34.00	53.39
$R_r\%$	86.52	43.64	12.93	47.69
$R_h\%$	75.28	-47.27	-15.52	4.16

Table 2. Results of processing one song of 60 sec length.

It seems save to say that it is very difficult, even for advanced listeners, to separate the proposed classes properly. However, there seems to be room for improvements in our system: (i) More excerpts for training data of variable playing styles and employing more powerful classifiers (e.g. SVMs) should definitely warrant an improvement. (ii) Misclassification of a component has a big impact on the overall accuracy since all the events it contains will be misclassified. Classifying individual onsets instead of components might reduce this impact. (iii) The NMF algorithm decomposes the magnitude spectrogram into constant source spectra that vary in gain over time, this is why it is so suitable for representing percussive instruments, because their spectrum doesn't change over time. The problem is that if pitched instruments are present in the mix, each note played is modelled as one component, resulting in a huge number of components. When too little components are chosen (as it is very often the case because of performance issues), pitched instruments are likely to be mixed within the components. While annotating the training data we found out that HH events are most likely to be affected by this phenomenon. HH events are the most frequent of the drum events in our evaluation song (44.62 %). That is why component number estimation would surely provide a major improvement.

4 REFERENCES

[1] S. Dixon. Onset detection revisited. In *Proc. of the DAFx*, pages 133–137, Montreal, Quebec, Canada, Sept. 18–20, 2006.

[2] T. Heittola and A. Klapuri. Locating segments with drums in music signals. In *Proc. of the 3rd ISMIR*, France, October 2002.

[3] M. Helen and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. of the 13th EUSIPCO*, Antalya, Turkey, September 2005.

[4] J. Paulus and T. Virtanen. Drum transcription with non-negative spectrogram factorisation. In *Proc. of the 13th EUSIPCO*, Antalya, Turkey, September 2005.

[5] K. Tanghe, S. Degroeve, and B. De Baets. An algorithm for detecting and labeling drum events in polyphonic music. In *Proc. of the first MIREX*, London, UK, September 11-15 2005.

[6] C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. of the 4th ICA*, Nara, Japan, April 2003.