DRUM DETECTION FROM POLYPHONIC AUDIO VIA DETAILED ANALYSIS OF THE TIME FREQUENCY DOMAIN

Christian Dittmar
Fraunhofer IDMT
Langewiesener Str. 22
98693 Ilmenau, Germany
dmr@idmt.fraunhofer.de

ABSTRACT

This publication presents a method for the automatic detection and classification of three distinct drum instruments in real world musical signals. The regarded instruments are kick, snare and hi-hat as agreed by the participants of the contest category Audio Drum Detection within the 2nd Annual Music Information Retrieval Evaluation eXchange (MIREX 2005). There are two challenging issues inherent to drum sound recognition in polyphonic music. The first problem is that the drum sound itself can vary greatly within the same instrument class, due to playing techniques, recording situation and electronic effects. The second apparent problem is the interference and masking with all other instruments sounding simultaneously with the drum in a musical signal, making it difficult to reliably detect occurrences of a certain drum type. The method outlined here achieves a solution to these problems by extending a source separation approach described in earlier publications with spectrogram templates and a more elaborate classification approach. Performance results of the system are given by the outcomes of the Audio Drum Detection contest within the MIREX 2005.

Keywords: ISMIR, MIREX, Drum Detection.

1 INTRODUCTION

For an in depth introduction to the motivation, problems and particularities of the audio drum detection task this document refers to [1]. Said paper also features a quite extensive section concerned with the state of the art at time of its publishing. For the sake of completeness this document pays attention to more recent works. In [2] Hidden Markov Models are compared against Support Vector Machines regarding their performance for the task of feature-based drum loop transcription. In [3] non-negative matrix factorization of the spectrogram is used in conjunction with pre-trained instrument spectra. Degradation of the model's performance is reported for signals that do not contain the expected instruments. To identify drum sounds in excerpts taken from polyphonic audio signals, feature-based percussion instrument sound models specialized on individual polyphonic audio recordings are proposed in [4] to achieve robustness against the distortion of features by concurrent instrument sounds. In [5] the signal is decomposed on the assumption on non-negative sparse source spectra and used for detection of kick and snare. A spectrogram-based template adaption and matching method is introduced with promising results for detection of kick and snare in [6].

2 SYSTEM OVERVIEW

2.1 Block diagram

An overview of the proposed system is presented in figure 1. It depicts that the signal processing chain can be subdivided into three main stages. The first step is the collection of onset times and corresponding onset spectra. Subsequently higher order statistical computations follow in order to estimate frequency and amplitude bases of the involved drum instruments, thus providing a decomposition into source components. Finally, a refinement stage that validates and enhances the intermediate results found so far delivers the classification and detection results. The subsequent sections will give a more in depth account of the different stages endorsed in the given diagram.

2.2 Onset Spectra Detection & Storage

The digital audio signals used for further analysis are mono files with 16 bit per sample at a sampling frequency of 44.1 kHz. A spectral representation of the pre-processed time signal is computed using a Short Time Fourier Transformation (STFT). Thereby a relatively large block-size in conjunction with high overlap is applied. The apodization function of choice is a Hann window. One could argue that the small hop-size implied by the overlap factor can not compensate for the time smearing introduced by the large window-size, but it can be shown to be useful for the subsequent processing steps. Based on the above mentioned steps a spectrogram representation of the original signal is derived. The unwrapped phase-information $\mathbf{\Phi}$ and the absolute spectrogram values $\mathbf{X}$ are taken into further consideration. The magnitude spectrogram $\mathbf{X}$ possesses n frequency bins and m frames. The time-variant slopes of each spectral bin are differentiated over all frames in order to decimate the influence of sustained sounds and to simplify the subsequent detection of transients. Half-wave rectification is applied in order to remove negative values introduced by the differentiation. This way, a non-negative difference-spectrogram $\hat{\mathbf{X}}$ is computed for the further processing.

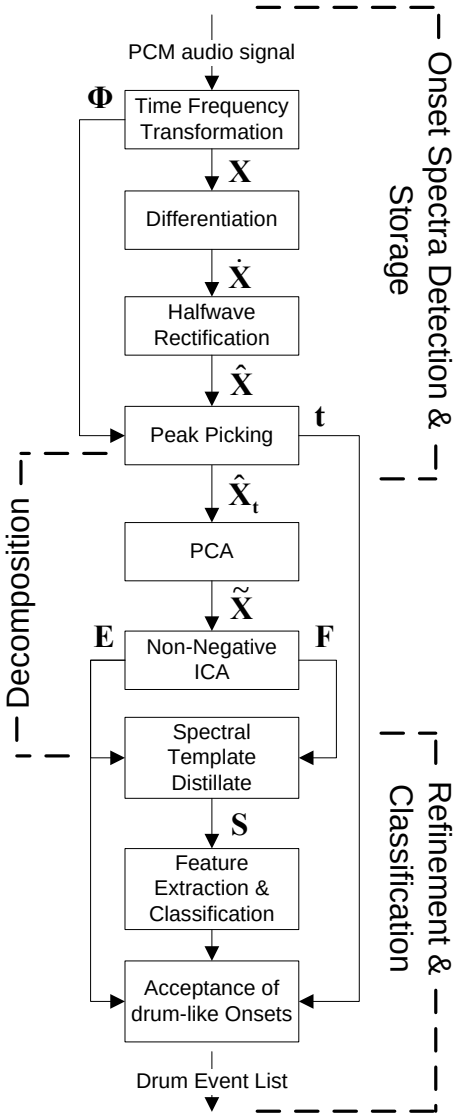

Figure 1. System Overview.

The detection of multiple local maxima positions $\mathbf{t}$ associated with transient onset events in the musical signal is conducted by means of peak picking in a suitable detection function. This function is derived from integrating over all bins of the non-negative difference spectrogram and smoothing the resulting vector. Two verification instances aim at avoiding the acceptance of small ripples for onsets. First, a time tolerance of 68 ms is defined which must at least occur between two consecutive onsets. Second, the unwrapped phase information of the original spectrogram serves as reliability function in this context. It can be observed that a significant positive phase jump must occur near the hypothetic onset-time $\mathbf{t}$. The main concept of the further process is the storage of one spectrum frame of the difference-spectrogram $\hat{\mathbf{X}}$ at the time of the onset. At this point, usage of an unconventionally large overlap is justified. The chance to capture the exact onset instant, where the characteristic attack phase is centred in the FFT window, increases drastically when applying a small hop-size. This way the precision of the onset spectra detection is increased and the common textbook statement concerning redundant overlap data is annulled. From the manifold of collected difference-spectrogram frames the significant spectral profiles related to the involved instruments will be gathered in the next stages.

2.3 Decomposition

From the steps described in the preceding section the information about the onset times $\mathbf{t}$ as well as the corresponding onset spectra $\hat{\mathbf{X}}_t$ is deduced. With regard to the goal of finding only a few significant subspaces, Principal Component Analysis (PCA) is applied to $\hat{\mathbf{X}}_t$. Using this well known technique it is possible to break down the whole set of collected spectra to a limited number of decorrelated principal components, thus resulting in a good representation of the original data with small reconstruction error. For this purpose an Eigenvalue Decomposition (EVD) of the dataset's covariance matrix is computed. From the set of eigenvectors the ones related to the d largest eigenvalues are chosen to provide the coefficients for a linear combination of the original vectors according to equation (1).

$$\tilde{\mathbf{X}} = \hat{\mathbf{X}}_t \cdot \mathbf{T} \qquad (1)$$

Thereby, $\mathbf{T}$ describes a transformation matrix which is actually a subset of the manifold of the covariance matrix eigenvectors. Additionally the reciprocal values of the eigenvalues are incorporated as scaling factors yielding not only a decorrelation but also a variance normalization, which in turn implies whitening [7]. Alternatively a Singular Value Decomposition (SVD) of $\hat{\mathbf{X}}_t$ according to [8] can achieve the same goal. With small modifications it is proven to be equivalent to the PCA using EVD [9]. The whitened components $\tilde{\mathbf{X}}$ are subsequently used as input for the Non-Negative Independent Component Analysis (ICA) in order to acquire the original source components. Non-Negative ICA uses the very intuitive concept of optimising a cost function describing the non-negativity of the components [10]. This cost function is related to the reconstruction error introduced by axis pair rotations of two or more variables in the positive quadrant of the joint probability density function (PDF). The assumptions for this model imply that the original source signals are positive and well grounded, which means they exhibit a non-zero PDF at zero, and they are to some extent linearly independent. The first concept is always fulfilled for the data considered in this publication, because the vectors subjected to ICA originate from the differentiated and half-wave rectified version $\hat{\mathbf{X}}$ of the amplitude-spectrogram $\mathbf{X}$, which does not contain any values lower than zero, but certainly some values at zero. The second constraint is taken into account when the spectra collected at onset times are regarded as linear superposition of a small set of original source-spectra characterizing the involved instruments. This seems, of course, to be a rather coarse approximation, but it holds up well in the majority of the

cases, which allows to separate the whitened components $\tilde{\mathbf{X}}$ into their potential sources $\mathbf{F}$ according to (2).

$$\mathbf{F} = \mathbf{A} \cdot \tilde{\mathbf{X}} \qquad (2)$$

Where $\mathbf{A}$ denotes the $d \times d$ unmixing matrix estimated by the ICA-process, which does actually separate the individual components $\tilde{\mathbf{X}}$. The sources $\mathbf{F}$ will be named spectral profiles from here forth. They are used to extract the spectrograms amplitude basis, hereafter referred to as amplitude envelopes according to (3).

$$\mathbf{E} = \mathbf{F} \cdot \mathbf{X} \qquad (3)$$

The extracted amplitude envelopes present relatively salient detection functions with sharp peaks, sometimes accompanied by smaller peaks and plateaus stemming from crosstalk effects. A percussiveness criterion [8] is computed on the amplitude envelopes to circumvent further inspection of irrelevant (non-drum like) components. Drum-like onsets are detected in the valid amplitude envelopes using conventional peak picking methods. Only peaks near the original times $\mathbf{t}$ are regarded as candidates. The value of the amplitude envelope's magnitude is assigned to every onset candidate at its position. If this value does not exceed a certain dynamic threshold then the onset is not accepted. The threshold varies over time according to the amount of energy in a larger area surrounding the onsets. Most of the crosstalk influences of harmonic sustained instruments as well as concurrent percussive instruments can be reduced in this step.

2.4 Refinement & Classification

Using the information about the probable onset times a spectrogram template is extracted for every valid component. This decisive refinement step is inspired by [6]. It can be achieved by peering through excerpts of the original spectrogram $\mathbf{X}$ near the times corresponding to the detection function's highest local maxima. The original spectrogram is taken into account to obtain multiple observations of the instruments in the time-frequency domain (preliminary templates), from which a statistically meaningful distillate of the actual instrument's spectrogram can be derived. Figure 2 depicts details of this process. The example given here shows a snap-shot amid the extraction of a snare drum template. The depiction is zoomed to a frequency range covering the first 200 bins. It can be seen that a suitably smoothed and scaled version of the corresponding spectral template is incorporated to trim interfering spectral peaks stemming from harmonic sustained instruments. This is achieved by element-wise minimum computation. The trimming operation is repeated for every desired time frame of the spectrogram template by moving forward both in $\mathbf{X}$ and the template.

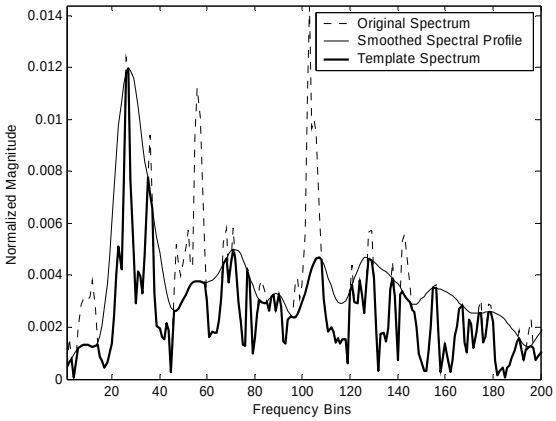

Figure 2. Details of the template extraction.

Multiple instances of such templates are extracted at different onset times ordered descending by the magnitude of the peaks in the corresponding amplitude envelope. This is motivated by the assumption, that the original instrument is probably present at those points with high amplitude. By virtually stacking all preliminary templates on top of each other and again performing element-wise minimum computation it is possible to capture the final spectrogram template $\mathbf{S}$ per instrument. It turned out, that a relatively small number of template observations is sufficient to distillate $\mathbf{S}$. Thus it is not necessary to sweep through whole songs in order adapt a template as proposed in [6]. The template represents the main characteristics of the detected drum instrument. It exhibits minimal interference of other instruments playing simultaneously and tends to smooth out spectral variance caused by slight playing variations of the drums. Figure 3 shows the comparison of an automatically extracted snare template (left) with an excerpt of the original spectrogram (right) where the snare sound is singly available at the beginning of this particular song. Although the single spectrogram was intentionally excluded from the template extraction process they still look very similar.

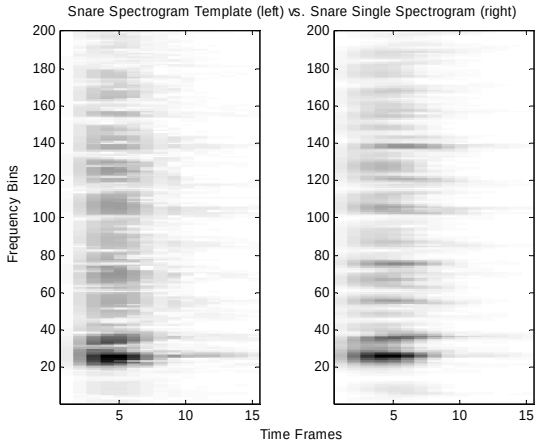

Figure 3. Comparison of spectrogram template and actual instrument spectrogram.

Because of the described characteristics the time-frequency templates are suited for feature extraction and classification of the drum sounds. Despite the fact that a time-frequency template offers the possibility to keep track of the development of certain spectral characteristics over time (e.g. when the drum sound is decaying) it turned out that a smoothed version of the mean spectrum over all template frames is distinctive enough for a separation into three classes.

Table 1. Features used for drum classification.

Feature Name	Feature dimension
Critical Band Energy	24
Spectral Centroid	1
Spectral Spread	1
Spectral Skewness	1
Spectral Kurtosis	1
Spectral Maximum Position	1

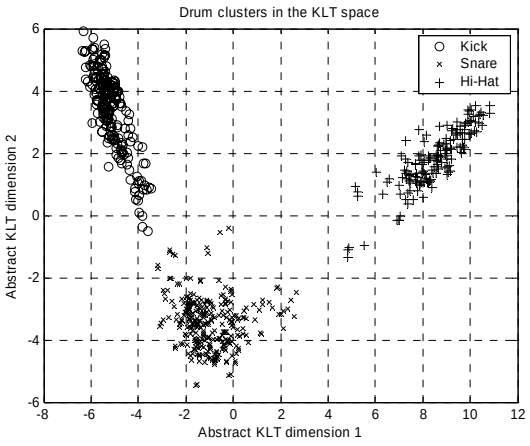

Figure 4. Distribution of the three drum classes in the KLT feature space.

The classification itself is based on a linear combination of the features given in table 1. The optimal feature combination is thereby predetermined in the sense of linear discriminant analysis. This minimization of intra-class variance and maximization of inter-class variance is achieved by Karhunen Loève Transformation (KLT) [9] applied to a normalized feature database extracted from a huge collection of single drum sounds with class labels. Figure 4 shows the distribution of the three training classes in the resulting two-dimensional KLT space. It can be seen that the clusters are clearly separated and the number of outliers is small. Mapping of the detected drum instruments to the three predefined classes is finally achieved by simple nearest neighbor classification using Euclidean distance measure in the KLT space. Concluding the process all valid onsets assigned to the detected instruments are stored in a text file using the format specified by the MIREX committee.

3 RESULTS

The performance assessment of the presented method shall be given by brief discussion of the respective MIREX results. Within the contest, approximately 50 files of both live and sequenced music were used as testbed. Many genres and various degrees of drum density (with regard to instrumentation as well as intensity) are encompassed by the files. Three collections of music were used, table 2 shows the average durations of the sound files.

Table 2. Information on the distinct test collections.

Submitted by	Average duration
Christian Dittmar	30 sec
Koen Tanghe	30 sec
Masataka Goto	4 min

To quantify the performance of each algorithm the participants agreed to use the F-measure (harmonic mean of the recall rate and the precision rate) for each of the three drum types, resulting in three F-measure scores and their average score. All participating algorithms were evaluated against music from each individual audio file collection. Subsequently the three collection scores were averaged to produce a composite score in which the method presented in this document positioned itself at rank five. It was outperformed by the winning algorithm submitted by Kazuyoshi Yoshii and three algorithmic variations presented by Koen Tanghe. For further detailed information please refer to [11]. It should be noted, that the current algorithm (handed in as a windows binary) could score a top result amongst its competitors in the category of processing speed. The duration of the complete run through all provided test files amounts to 673 seconds on an Intel P4 machine with 3 GHz processor and 3 GB of RAM. Unfortunately, this number is not directly comparable to the other participants due to the usage of distinct computers and frameworks.

4 CONCLUSIONS

The method outlined here proves to yield quite acceptable performance results in comparison to its competitors. It is the strong belief of the author that positive synergy effects can be achieved if concepts of the winning algorithms in this particular MIREX contest category are incorporated into the system. It would, for example, be very interesting to see how the template matching principle described in [6] works together with the spectrogram templates extracted according to this document.

5 ACKNOWLEDGEMENTS

The author would like to express his gratitude to the whole MIREX staff for their combined efforts to make this forward-looking contest possible. Special thanks go to Christian Sailer, Stefan Fuhrmann and Christian Uhle for proofreading of this document and valuable suggestions to its clarity.

REFERENCES

[1] C. Dittmar, C. Uhle, “Further Steps towards Drum Transcription of Polyphonic Music”, in Proc. of the AES 116th Convention, 2004.

[2] O. Gillet, G. Richard, “Automatic Transcription of Drum Loops”, in Proc. Int. Acoustics, Speech, and Signal Processing Conference (ICASSP), 2004.

[3] J. Paulus, A. Klapuri, “Drum Transcription with Non-Negative Spectrogram Factorisation”, to be appearing in Proc. European Signal Processing Conference (EUSIPCO), 2005.

[4] V. Sandvold, F. Gouyon, P. Herrera, “Percussion classification in polyphonic audio recordings using localized sound models”, in Proc. Int. Music Information Retrieval Conference (ISMIR), 2004.

[5] T. Virtanen, “Sound Source Separation using Sparse Coding with Temporal Continuity Objective”, in Proc. Int. Computer Music Conference (ICMC), 2003.

[6] K. Yoshii, M. Goto, H.G. Okuno, “Automatic Drum Sound Description for Real-World Music using Template Adaption and Matching Methods”, in Proc. Int. Music Information Retrieval Conference (ISMIR), 2004.

[7] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley & Sons, 2001.

[8] C. Uhle, C. Dittmar, T. Sporer, “Extraction of Drum Tracks from polyphonic Music using Independent Subspace Analysis”, in Proc. of the Fourth International Symposium on Independent Component Analysis, 2003.

[9] A.Webb, Statistical Pattern Recognition, Wiley & Sons, 2002.

[10] M. Plumbley, “Algorithms for Non-Negative Independent Component Analysis”, in IEEE Transactions on Neural Networks, 14 (3), 2003.

[11] J.S. Downie, 2005 MIREX Contest Results - Audio Drum Detection, http://www.music-ir.org/evaluation/mirex-results/audio-drum/index.html