Subband-based Drum Transcription for Audio Signals

George Tzanetakis Ajay Kapur Richard I. McWalter
University of Victoria
3800 Finnerty Rd.
Victoria, B.C., Canada
Email: gtzan@cs.uvic.ca, ajay@ece.uvic.ca, rim@uvic.ca

***Abstract*— Content-based analysis of music can help manage the increasing amounts of music information available digitally and is becoming an important part of multimedia research. The use of drums and percussive sounds is pervasive to popular and world music. In this paper we describe an automatic system for detecting and transcribing low and medium-high frequency drum events from audio signals. Two different subband front-ends are utilized. The first is based on bandpass filters and the second is based on wavelet analysis. Experimental results utilizing music, drum loops and Indian tabla thekas as signals are provided. The proposed system can be used as a preprocessing step for rhythm-based music classification and retrieval. In addition it can be used for pedagogical purposes.**

I. Introduction

The amount of music digitally available is exploding. Popular portable music players can hold thousands songs and online distribution of musical content is reality. Content-based analysis of audio and video data is one of the major challenges of multimedia research. Music Information Retrieval (MIR) is an emerging area of multimedia research that explores algorithms and tools for analyzing large collections of music for retrieval and browsing. Rhythmic structure is an essential component in representing and understanding music.

This paper explores the use of signal processing techniques to provide rhythmic transcriptions of polyphonic music and drum patterns. Transcription refers to the process of converting the audio recording to a symbolic representation similar to a musical score. The transcribed symbolic representation can then be used for classification and retrieval tasks.

In addition, it can be rendered using synthetic sound. That way for example a student drummer can isolate the drum accompaniment of a song from the rest of the music. Similarly, a student of tabla (Indian drums) could use a drum transcription system to slow down and learn a fast complicated rhythm recorded by their guru (master). Another possibility is a disk jockey (DJ) tapping out a particular drum pattern and retrieving all songs which have similar structure. These are some of the representative scenarios motivating this work.

The remainder of the paper is structured as follows: related work is described in section 2. Section 3 provides a description of the system and Section 4 provides experimental results. Some directions for future work are provided in Section 5.

II. Related Work

Most existing work in rhythm-based music information retrieval has concentrated on tempo induction and classification of individual drum sounds. An early representative example of a tempo induction system is [1]. More recently there has been some work in drum sound transcription that is more relevant to this work. A system for *Query-by-Rhythm* was introduced in [2]. Rhythm is stored as strings turning song retrieval into a string matching problem. The authors propose an L-tree data structure for efficient matching. The similarity of rhythmic patterns using a dynamic programming approach is explored in [3]. A system for the automatic description of drum sounds using a template adaption method is [4]. A system for the classification of dance music based on periodicity patterns is presented in [5]. A system for classifying percussive sounds in audio signals is proposed in [6] and [7] describes the extraction of drum patterns and their description using the MPEG7 framework for multimedia content description. A transcription system for drum loops using Hidden Markov Models and Support Vector Machines is described in [8].

III. System Description

A. Overview

In order to extract information for transcription and retrieval applications we perform what we term “boom-chick” analysis. The idea is to detect the onset of low frequency events typically corresponding to bass drum hits and high frequency events typically corresponding to snare drum hits from the polyphonic audio sources. This is accomplished by onset detection using adaptive thresholding and peak picking on the amplitude envelope of two of the frequency bands. Unlike many of the proposed approaches in drum transcription we do not utilize a classification model because that would constrain the application of our method to specific types of sounds. Even when a classification approach is utilized a data-driven front end as the one described in this work can be used as a preprocessing step for training. In this paper two front-ends are compared. The first one is based on using regular bandpass filters designed for detecting drum sounds. The second one is based on the wavelet transform and is similar to the Beat Histogram calculation front-end described in [9]. Experiments in section IV compare results using the two front-ends.

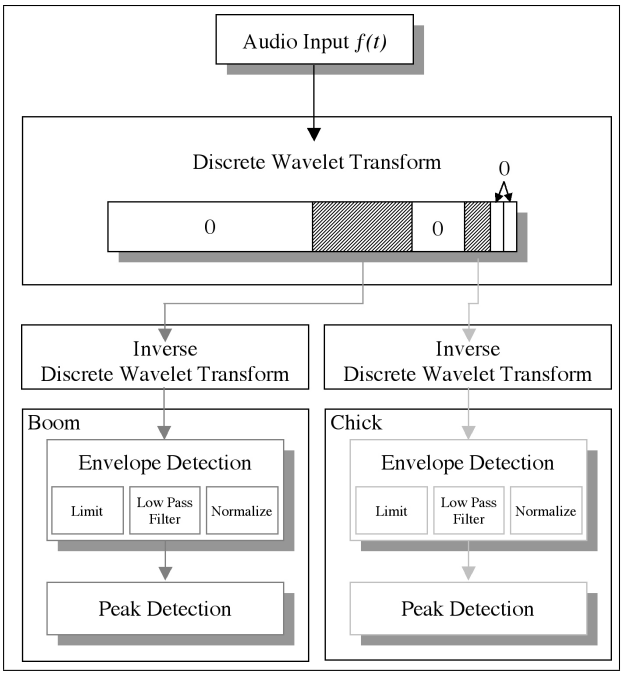

Fig. 1. Wavelet Front-End for drum sound detection

B. Filter analysis

In order to detect drum sounds two subbands ("boom" and "chick") are utilized. The signal is analyzed in windows of approximately 3 seconds. The frequency range for the low subband is 30Hz to 280Hz. This was determined empirically to give good results over a variety of drum loops and music recordings. A simple bandpass filter implemented in the frequency domain was used to select the "boom" band. The frequency range for the high subband is 2700 Hz- 5500 KHz. The "chick" subband is implemented using a Butterworth filter.

C. Wavelet-based analysis

The second front-end is based on decomposing the signal into different frequency bands using a Discrete Wavelet Transform (DWT) similarly to the method described in [9] for the calculation of *Beat Histograms*. Figure 1 shows how a window of an audio signal (approximately 3 seconds) is converted to the wavelet domain with a fourth order Daubechies wavelet. The "boom" band and the "chick" band were determined through experimentation although unlike the filter-based approach the boundaries were constrained by the octave characteristics of the dyadic wavelet transform. The bands with the most information were approximately 300Hz-600Hz for the "boom" and 2.7KHz-5.5KHz for the "chick". To convert the signals back to the time domain, the wavelet coefficients for all the bands except the chosen one are set to zero and then the Inverse Wavelt Transform. That way the resulting subband signal is in the time domain.

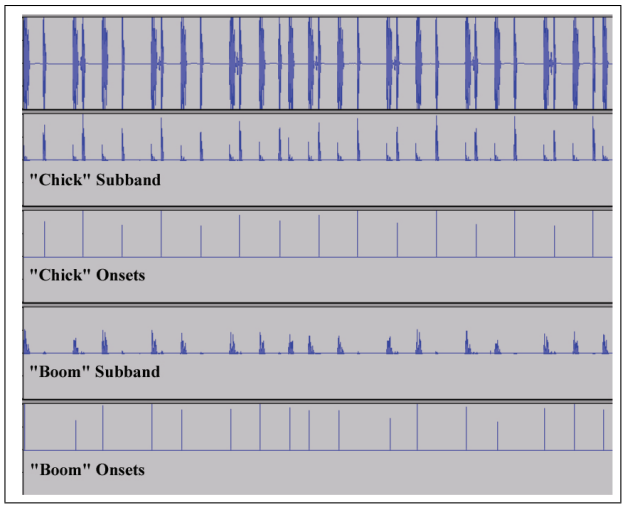

Fig. 2. Audio signal "Boom-Chick" decomposition

D. Envelope extraction and Onset Detection

Once the subbands are calculated using either front-end, they are subsequently processed to find the onset of the "boom" and "chick" sounds. First the envelope of each subband is calculated by using full wave rectification followed by low pass filtering and normalization. Once the envelope is extracted an adaptive peak detection algorithm based on thresholding is utized to find the onset times of the percussive sounds. If the spacing between adjecent peaks is small, then only the highest one will be selected. Figure 2 shows how a drum loop can be decomposed into "boom" and "chick" bands and the corresponding detected onsets.

E. Transcription

The goal of transcription is to convert the two sequences of onset times for the "boom" and "chick" bands into symbolic form. Music notation is relative to the tempo of the piece which means that two pieces that have the same drum pattern played faster and slower will still have the same notation. The first step for transcription is to calculate the IOI (Inter-onset Intervals) which are the time differences in samples between onset positions. The IOIs are subsequently quantized in order to ignore small variations in tempo. A combination of heuristics based on music knowledge and clustering of the quantized IOIs is used to select the IOI that corresponds to a quarter note. Once this basic rhythmic unit is established all IOIs are expressed relative to it as integer ratios. The resulting ratios can then be directly rendered in music notation. Currently the output of the system is a textual representation of the durations. Figures 3, 4 show a common music notation and an Indian tabla notation rendering of the output of the system. Even though the graphic noatation was rendered manually it corresponds directly to the output of the drum transcription system.

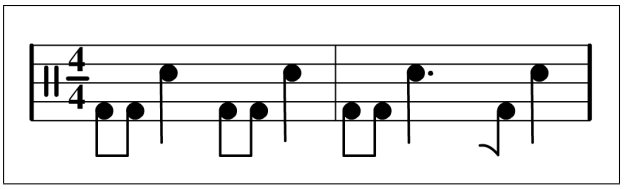

Fig. 3. Transcribed drum loop (bass and snare drum)

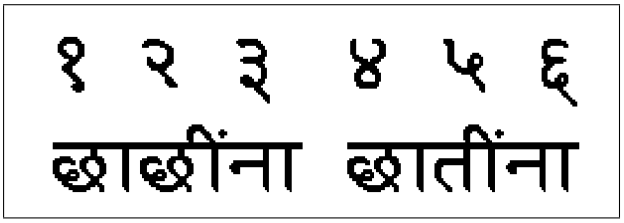

Fig. 4. Transcribed Indian tabla theka (Dadra - 6 beats)

IV. Experiments

A. Experimental Setup

Three sample data sets were collected and utilized. They consist of techno beats, tabla *thekas* and music clips. The techno beats and tabla *thekas* were recorded using DigiDesign Digi 002 ProTools at a sampling rate of 44100 Hz. The techno beats were gathered from Dr. Rex in Propellerheads Reason. Four styles (Dub, House, Rhythm & Blues, Drum & Bass) were recorded (10 each) at a tempo of 120 BPM. The tabla beats were recorded with a pair of AKG C1000s to obtain stereo separation of the different drums. Ten of each of four "thekas" (meaning beats per cycle) were recorded (Tin Taal Theka (16), Jhaap Taal Theka (10), Rupak Theka (7), Dadra Theka (6)). The music clips consist of jazz, funk, pop/rock and dance music with strong rhythm. The system has been implemented using the MARSYAS [1] which is a software framework for prototyping audio analysis and synthesis applications.

B. Results

The evaluation of the system was performed by comparative testing between the actual and detected beats by two drummers. After listening to each track, false positive and false negative drum hits were detected seperately for each type ("boom" and "chick"). False positives are the set of instances in which a drum hit was detected but did not actually occur in the original recording. False negatives are the set of instances where a drum hit occurs in the original recording but is not detected automatically by the system. In order to determine consistency in annotation, five random samples from each dataset were analyzed by both drummers. The results were found to be consistent and therefore the ground-truth annotation task was split evenly among the two drummers. These two expert users also provided feedback for the fine tuning of the system.

[1] `http://marsyas.sourceforge.net`

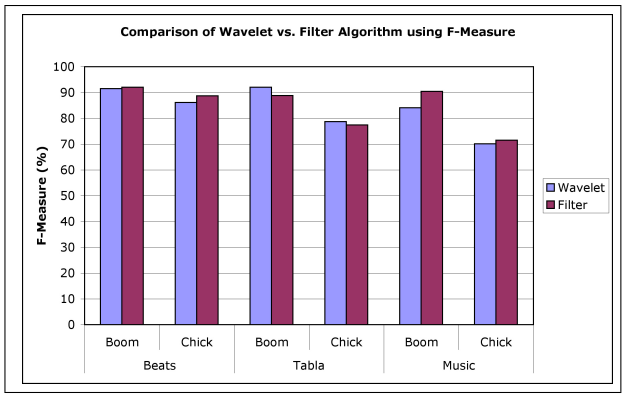

Fig. 5. Comparison of filter and wavelet frontend

The results are summarized using the standard precision and recall measures. Precision measures the effectiveness of the algorithm by dividing the number of correctly detected hits (true positives) by the total number of detected hits (true positives + false positives). Recall represents the accuracy of the algorithm by dividing the number of correctly detected hits (true positives) by the total number of actual hits in the original recording (false negatives + true positive). Recall can be improved by lowering precision and vice versa. A common way to combine these two measures is the so called F-measure defined as (P is precision, R is recall and higher values of the F-measure indicate better retrieval performance):

$$F = \frac{2 * P * R}{P + R} \tag{1}$$

Figure 5 summarizes the detection results using the two front-ends for the 3 different audio data collections. As can be seen from the figure the dection of the low frequency "boom" events is overall better than the "chick" events. This is expected as there is less variability in bass drum sounds and less interference from other instruments and percussive sounds. The results are better for the drum loops and tabla thekas where there are only percussive sounds. As expected the results are not as good for the polyphonic audio clips where the presence of other interferring sounds, such as singers, guitars, and other instruments, makes detection harder. The difference between the feature front-ends is statistically insignificant given the size of the collections used. As the filter front-end is faster to compute, we provide more detailed results for that front-end in Tables I and II. Drum transcription of a 30-second clip on a Pentium III 1.2 GHz takes approximately 8 seconds for the wavelet front-end and 2 seconds for the filter front-end. These performance measurements are relative as there was no effort to optimize the run-time performance of the system. Also the analysis is causal (requires a single pass) and therefore can be performed on streaming audio.

TABLE I

"CHICK" HIT DETECTION RESULTS FOR FILTER FRONTEND

Category	Recall	Precision	F-measure
Rnb	0.94	0.97	0.95
Dnb	0.83	0.86	0.84
Dub	0.92	0.84	0.88
Hse	0.93	0.83	0.88
Average	0.90	0.87	0.89
Dadra	0.53	1.00	0.69
Rupak	0.51	1.00	0.68
Jhaptaal	0.79	1.00	0.88
Tintaal	0.73	1.00	0.84
Average	0.64	1.00	0.77
Various	0.60	0.51	0.55
Dance	0.94	0.63	0.75
Funk	0.93	0.58	0.72
Average	0.82	0.57	0.67

TABLE II

"BOOM" HIT DETECTION RESULTS FOR FILTER FRONTEND

Category	Recall	Precision	F-measure
Rnb	0.85	0.83	0.84
Dnb	0.93	0.88	0.90
Dub	0.92	0.97	0.94
Hse	0.99	1.00	0.99
Average	0.92	0.92	0.92
Dadra	0.98	0.90	0.94
Rupak	1.000	0.56	0.72
Jhaptaal	0.98	0.93	0.95
Tintaal	0.91	0.97	0.94
Average	0.97	0.84	0.89
Various	0.88	0.79	0.83
Dance	0.92	0.89	0.90
Funk	0.86	0.95	0.90
Average	0.89	0.88	0.88

V. FUTURE WORK

There are many directions for future research. We are currently experimenting with the use of the detected onsets to train song-specific models for drum sounds using machine learning techniques. Once those sounds have been detected they can be used to improve the results by removing false positives. This is similar to the template adaption approach described in [4].

Another direction we are investigating is the use of MPEG7 descriptors for drum sound classification. Rhythmic information has been shown to be an important factor for genre and style classification. The drum transcription system described in this paper is currently used to build a retrieval and classification system for DJs and radio producers. Another possibility is the use of the proposed drum transcription system to transcribe vocal percussion as described in [10]. Finally, we are exploring the use of *Lilypond* [2], which is a music engraving program, to produce directly drum notation from audio signals. Indian tabla notation can also be directly rendered using Hindi fonts.

[2]`http://www.lilypond.org`

One of the difficulties of drum transcription is the availiability of ground truth. Using expert listeners as was done in this paper is difficult for large data sets. We are exploring the use of sensor-enhanced drums to directly obtain ground truth data from drum performances.

ACKNOWLEDGMENT

The authors would like to thank the student of the first CSC 484 Music Information Retrieval course at University of Victoria for their ideas, support and criticism during the development of this project. Thanks to Peter Driessen and Andrew Schloss for their support.

REFERENCES

[1] E. Schierer, "Tempo and beat analysis of acoustic musical signals," *Journal Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[2] J. Chen and A. Chen, "Query by rhythm: An approach for song retrieval in music databases," in *Proc. Int. Workshop on Research Issues in Data Engineering*, 1998.

[3] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. Inter. Conf. on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[4] K. Yoshii, M. Goto, and H. Okuno, "Automatic drum sound description for real world music using template adaption and matching methods," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2004.

[5] S. Dixon, E. Pampalk, and G. Widmer, "Classification of dance music by periodicity patterns," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, Baltimore, USA, 2003.

[6] F. Gouyon, F. Pachet, and O. Delerue, "On the use of zero-crossing rate for an application of classification of percussive sounds," in *Proc. of Int. Conf. on Digital Audio Effects*, Verona, Italy, 2000.

[7] M. Gruhre, C. Uhle, C. Dittmar, and M. Cremer, "Extraction of drum patterns and their description within the mpeg-7 high-level framework," in *Proc. Inter. Conf. on Music Information Retrieval(ISMIR)*, Barcelona, Spain, 2004.

[8] O. Gillet and G. Richard, "Automatic transcription of drum loops," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, 2004.

[9] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, July 2002.

[10] A. Kapur, M. Benning, and G. Tzanetakis, "Query-by-beatboxing: Music information retrieval for the dj," in *Proc. Inter. Conf. on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004.