Locating Segments with Drums in Music Signals

Toni Heittola
Tampere University of Technology
P.O.Box 553
FIN-33101 Tampere, Finland
+358 3 3115 3788
toni.heittola@tut.fi

Anssi Klapuri
Tampere University of Technology
P.O.Box 553
FIN-33101 Tampere, Finland
+358 3 3115 2124
klap@cs.tut.fi

ABSTRACT

A system is described which segments musical signals according to the presence or absence of drum instruments. Two different yet approximately equally accurate approaches were taken to solve the problem. The first is based on periodicity detection in the amplitude envelopes of the signal at subbands. The band-wise periodicity estimates are aggregated into a summary autocorrelation function, the characteristics of which reveal the drums. The other mechanism applies straightforward acoustic pattern recognition with mel-frequency cepstrum coefficients as features and a Gaussian mixture model classifier. The integrated system achieves 88 % correct segmentation over a database of 28 hours of music from different musical genres. For the both methods, errors occur for borderline cases with soft percussive-like drum accompaniment, or transient-like instrumentation without drums.

1. INTRODUCTION

The presence/absence of drum instruments is an important high-level descriptor for music classification and retrieval. In many cases, exactly expressible descriptors are more efficient for information retrieval than more ambiguous concepts such as musical genre. Information about the drums can also be used in audio editing, or in further analysis, e.g. in music transcription, metrical analysis, or rhythm recognition.

The aim of this paper is to present a drum detection system, which would be as generic as possible. The problem of drum detection in music is more difficult than what it seems at a first glance. For a major part of techno or rock/pop music, for example, detection is more or less trivial. However, a detection system designed for these musical genres does not generalize to the others. Music contains a lot of cases that are much more ambiguous. Drums go easily undetected in jazz/big band music, where only hihat or cymbals are softly played at the background. On the other hand, erroneous detections may pop up for pieces with acoustic steel-stringed guitar, pizzicato strings, cembalo, or staccato piano accompaniment, to mention some examples.

2. METHODS

Drum instruments in Western music typically have a clear stochastic noise component and they can be recognized based on that stochastic component [1]. A sinusoids+noise spectrum model was used to extract the stochastic parts of acoustic musical signals, because residual signal has significantly better "drums-vs-other" ratio than the input signal [2].

2.1 Periodicity Detection Approach

Periodicity is characteristic for musical rhythms. Drum events typically form a pattern which is repeated and varied over time. As a consequence, the time-varying power spectrum of the signal shows clear correlation with a time shift equal to the pattern length in the drum track. We propose that the presence of drums can be detected by measuring this correlation in musical signals. This evaluates an underlying hypothesis that periodicity of stochastic signal components is a universally characteristic of musical signals with drums. In order to alleviate the interference of other musical instruments, periodicity measurement is performed in the residual signal after preprocessing with a sinusoidal model.

Band energy ratio (BER) feature was used to model signals rough spectral energy distribution. BER is defined as the ratio of the energy at a certain frequency band to the total energy [3]. Since human auditory perception does not operate on a linear frequency scale, we apply a filter bank consisting of triangular filters spaced uniformly on the mel-scale. At each frequency band, an autocorrelation function (ACF) is calculated over the BER values within a three-second long sliding analysis window, intended to capture a few patterns of even the slowest rhythms. Despite the preprocessing, also other instruments cause peaks to the bandwise autocorrelation functions. Its effects are minimized with weighting bands differently before forming the summary autocorrelation function (SACF). The SACF will be finally mean-normalized to get real peaks step out better from the SACF [4]. Overview of whole system is shown in Figure 1

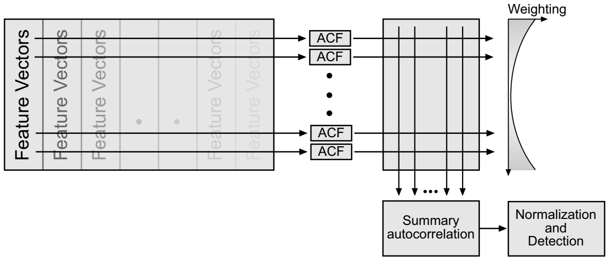

Figure 1. System overview.

As can be seen in Figure 2, periodic drum events produce also a periodic SACF. In order to robustly detect this, time-scaled (by factors 2 and 3) versions of SACF are added to the original SACF to yield enhanced SACF (ESACF). Thus peaks at integer multiples of a fundamental tempo are used to enhance the peaks of a slower tempo. This technique has been adopted from [5].

The region of interest in the ESACF is determined by reasonable tempo limits (35 beats per minute and 120 beats per minute). It should be noted that due to the above describe enhancement procedure, these limits actually corresponds to 35 and 360 in SACF. This wide tempo range is essential because the rate of playing certain drum instruments is typically an integer multiple of tempo, and causes a clear peak in the SACF. Final detection is carried out by measuring the absolute maximum value of ESACF within the given tempo limits.

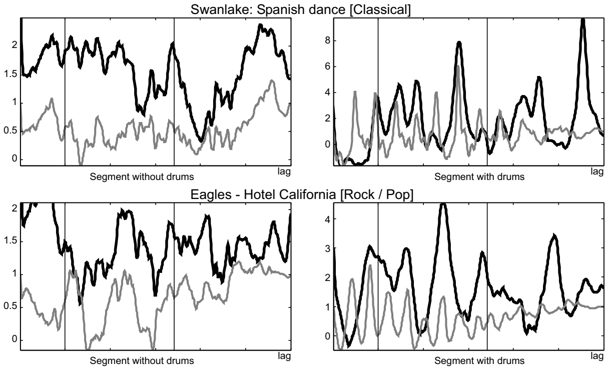

Figure 2. Representative SACFs (gray line) and ESACFs (black line) (Tempo limits marked in the plots).

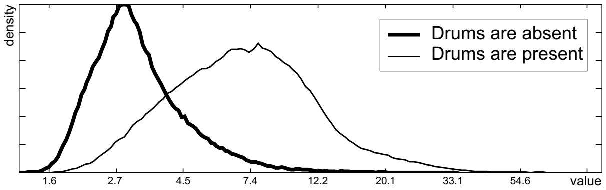

Figure 3. Unit area normalized feature value distributions.

2.2 Acoustic Pattern Recognition Approach

Motivated by characteristic spectral energy distributions of drum sounds, we studied the ability of Mel-frequency cepstral coefficients (MFCC) to indicate the presence of the drums in musical signals. We used 16 coefficients (both static and delta), calculated in 20ms frames with ¼ overlap, as features for a classifier. Two different classifiers were used, one based on Gaussian Mixture Models (GMM) and k-nearest neighbour classifier (k-NN) with Mahalanobis distance measure.

3. Simulations And Results

A database of 397 entire musical pieces from different genres was used to evaluate the two drum detection schemes. “Presence of drums” was defined to include the requirement that the drum is played in a rhythmic role. An individual piece was randomly selected either to the training set or the test set. All detection results are shown in Table 1.

For periodicity detection approach, training set was used to estimate the distribution of maximum values in ESACF within the given tempo limits (see Figure 3), and to determine a threshold value used in detection. The reason why the distributions of the two classes overlap rather much is that the stochastic residual contains harmonic components and beginning transients from other instruments, too, and in some cases these show very much drum-like periodicity. Thus the starting hypothesis that periodic stochastic components reveal drum events was still mainly right. More attention should be paid for the preprocessing system in order to make concluding remarks.

In order to perform classification with GMM training sets were used to estimate model parameters for the two classes, one model for music with drums and another for music without drums. Performance was slightly better than with the system periodicity detection approach, but performance was not evenly distributed within different musical styles. Although a high performance is obtained for one class (e.g. drums present), the other fails within the individual musical style. In other words, the system starts to recognize the musical style rather than the drums.

Table 1. Detection results. Table notation: <detection rate> (<rate for segments with drums> / <without drums>)

Genre[1]	Periodicity	GMM [2]	k-NN [3]
Classical (27%)	**83%** (84/ 78)	**90%** (97/ 39)	**83%** (88/44)
Electronic/Dance (7%)	**91%** (61/ 96)	**89%** (49/96)	**86%** (25/96)
Hip Hop/Rap (3%)	**87%** (70/ 88)	**94%** (26/ 98)	**95%** (11/99)
Jazz/Blues (16%)	**75%** (38/ 79)	**74%** (58/ 76)	**77%** (47/80)
Rock/Pop (29%)	**83%** (82/ 83)	**92%** (68/ 95)	**89%** (46/93)
Soul/RnB/Funk (11%)	**78%** (80/ 78)	**91%** (77/ 93)	**89%** (46/93)
World/Folk (7%)	**69%** (52/ 92)	**68%** (48/ 95)	**60%** (32/95)
Total	**81%** (77/83)	**87%** (84/88)	**83%** (71/89)

[1] Portion of evaluation database is put in brackets.

[2] GMM classification with two models. (MFCC + ΔMFCC, model order 24, three-second test excerpts)

[3] k-NN classification with *k=5* and three-second test excerpts.

Since two drum detection systems are based on different information, one would thus guess that the combination of the two systems would perform more reliably than either of them alone. But only minor improvement (1-2%) was achieved with integration (Periodicity + GMM). This is due to the fact that both of the systems typically misclassify within the same intervals. For example, jazz pieces where drums are played quite softly with brush, or ride cymbal is continually tapped are likely to be misclassified with both systems. In some cases, the misclassification might be acceptable, since the drums are difficult to detect even for a human listener.

4. SUMMARY AND CONCLUSIONS

Two different drum detection schemes were described and evaluated. The obtained results are rather close to each other and, somewhat surprisingly, the combination performs only slightly better. Achieved segmentation accuracy of the integrated system was 88 % over a database of varying musical genres. In order to construct a substantially more accurate system, it seems that more complicated sound separation and recognition mechanism would be required. In non-causal applications, longer analysis excerpts and the global context can be used to improve the performance.

5. REFERENCES

[1] Fletcher, N. H. and Rossing, T. D., The Physics of Musical Instruments. Springer-Verlag, New York, 1991.

[2] Virtanen, T., Audio signal modeling with sinusoids plus noise. Master’s thesis, Department of Information Technology, Tampere University Of Technology, 2000

[3] Peltonen,V., Computational Auditory Scene Recognition. In Proc. ICASSP, Orlando, Florida, May 2002.

[4] de Cheveigné, A. and Kawahara, H., YIN a fundamental estimator for speech and music. JASA, 2002.

[5] Tolonen, T. and Karjalainen, M., A computationally efficient multipitch analysis model, IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 6, Nov. 2000