COMBINING TEMPORAL AND SPECTRAL FEATURES IN HMM-BASED DRUM TRANSCRIPTION

Jouni Paulus, Anssi Klapuri
Institute of Signal Processing
Tampere University of Technology

ABSTRACT

To date several methods for transcribing drums from polyphonic music have been published. Majority of the features used in the transcription systems are "spectral": parameterising some property of the signal spectrum in a relatively short time frames. It has been shown that utilising narrow-band features describing long-term temporal evolution in conjunction with the more traditional features can improve the overall performance in speech recognition. We investigate similar utilisation of temporal features in addition to the HMM baseline. The effect of the proposed extension is evaluated with simulations on acoustic data, and the results suggest that temporal features do improve the result slightly. Demonstrational signals of the transcription results are available at http://www.cs.tut.fi/sgn/arg/paulus/demo/. [1]

1 INTRODUCTION

Systems for automatic transcription of music have gained a considerable amount of research effort during the last few years. From the point of view of music information retrieval, these can be considered as tools for rising from the acoustic signal to a higher level of abstraction that correlates better with the content of interest. Here we focus on the transcription of drums: locating and recognising sound events created by drum instruments in music.

Several methods have been proposed for drum transcription. Some of them are based on locating the onsets of prominent sound events, extracting a set of features from the locations of the onsets, and classifying the events using the features. Systems of this category include the method by Tanghe et al. [11] using support vector machines (SVMs) as classifiers, and a system using template adaptation and iterative musical pattern based error correction by Yoshii et al. [13]. As an extension to the systems relying only on acoustic data, Gillet and Richard have proposed a multi-modal system which uses also visual information of the drummer playing [4]. A system using hidden Markov models (HMMs) was presented in [8].

[1] This work was supported by the Academy of Finland, project No. 5213462 (Finnish centre of Excellence program 2006 - 2011).

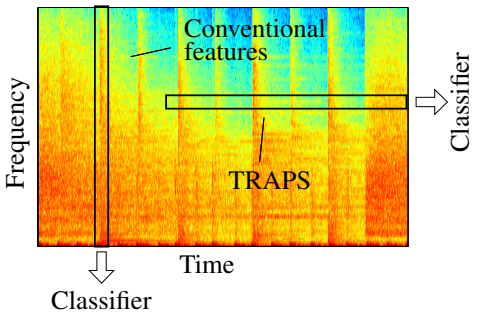

Figure 1. The basic idea of temporal features illustrated. Instead of short wide-band frames of data, features are calculated from long narrow-band frames. (After [6].)

In polyphonic music the presence of other instruments makes the transcription more difficult, since they are effectively noise for drum transcription systems. An alternative to the previous approaches is to try to separate the drums from polyphonic music, or to separate each drum to its own stream. The separation can be done blindly without any prior templates for the drums, as is done by Dittmar and Uhle [2], or by using a dictionary for different drums [9]. Several methods, both blind and dictionary-based, developed by FitzGerald et al. are detailed in [3]. For a more description of the earlier methods refer to [3].

Majority of the features used in the recognisers are "spectral": parameterising some property of the signal spectrum in a relatively short (e.g., 10 ms) time frames. Features describing the temporal evolution of the signal are usually limited to the first temporal derivatives of spectral features, and in essence they still are short-time features. Some systems hand the responsibility of modelling the temporal evolution of the features over to an HMM architecture: different states describe different time instants of the modelled event. Hermansky and Sharma proposed an alternative for this in [6] in the form of using TRAPS (TempoRAl PatternS) features describing energy evolution at different subbands. The main idea behind TRAPS is illustrated in Figure 1. They showed that utilising the information of how the energy evolves on several subbands in one-second frames it was possible to improve the performance of a baseline speech recogniser which used only short-time cepstral features.

Features from subband envelopes have been used earlier also in other music related applications, such as mu-

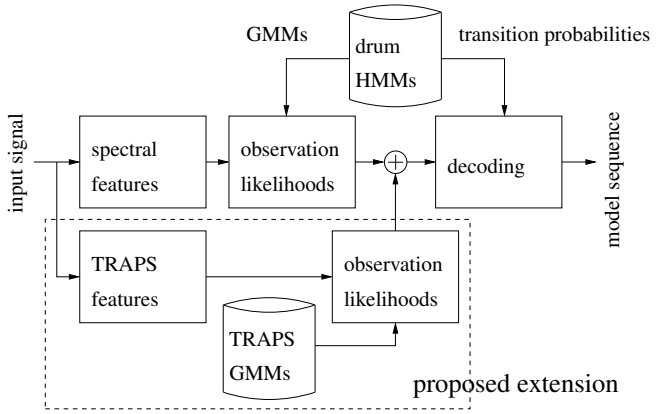

Figure 2. Block diagram of the full system with the proposed extension circled with dashed line.

sical piece structure analysis [10], genre classification [7], and automatic record reviews [12]. However, the features used in these works were concentrated on the modulations of the envelopes whereas we are interested in certain events: the drum hits.

We propose to utilise the information from temporal features in addition to the earlier HMM-based system [8], and show that they do increase the performance. Temporal features suit for drums, because drums are usually short events and do not have any "stable" state as e.g. harmonic sounds may have. The baseline HMM system is described in Section 2.1. The added temporal features are detailed in Section 2.2. Methods for combining the information from temporal features to the baseline system are described in Section 2.3. The performance of the resulting system is evaluated with simulations described in Section 3. Finally, conclusions are given in Section 4.

2 PROPOSED METHOD

The block diagram of the system including the proposed extension is illustrated in Figure 2. The baseline system extracts a set of spectral features from the input signal and estimates observation likelihoods for all HMM states using Gaussian mixture models (GMMs). Finally, the transcription is obtained by finding out the best state and model sequence to explain the observed features. The extension adopts the idea from [6] and assumes that temporal features can provide information which can correct some of the errors made by the baseline system. The information provided by the proposed extension is added to the baseline system in the observation likelihood stage before decoding.

2.1 Baseline HMM Recogniser

The baseline HMM system is the one published earlier in [8]. Each combination of the target drums is modelled with a HMM and one HMM serves as a background model for the situation when none of the target drums is playing. These models are combined into a network whose idea is illustrated in Figure 3. At each time frame the system is

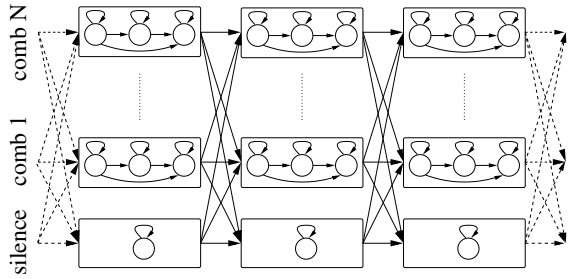

Figure 3. The idea of the used HMM model network consisting of drum combinations ("comb 1" and "comb N") and the silence model.

in one state of one of the combination models. After the system exits a combination model it may enter another combination or the background model. In the recognition phase, the best path through the models is searched using token passing algorithm [14].

When handling polyphonic music signals, the input is passed through a sinusoids+residual -modelling. The modelled sinusoids are subtracted from the original signal and the residual is regarded as the input signal for the further processing. It is assumed that some components of pitched instruments are modelled with the sinusoids and their contribution is reduced in the residual. As most of the sound generated by the drums is stochastic, the modelling does not affect them much. Then the signal is divided into 23.2 ms frames with 50% overlap and the following set of features is extracted: 13 mel-frequency cepstral coefficients and their first order time differences, energy of the signal, spectral centroid, kurtosis, spread, slope, flatness, and roll-off, and log-energy from 8 octave-spaced bands.

The used HMM architecture uses four states for all of the drum combinations and one state for the background model. The feature distributions in the states are modelled with GMMs with two components in states belonging to the drum models and 10 components in the background model. State transitions are allowed only to the state itself and to the next state.

2.2 Temporal Features

Hermansky et al. used one-second frames of subband energy envelopes sampled at 100 Hz as the input to a multilayer perceptron (MLP) classifier. Here we use only the main idea of temporal features and divert from the original method. Even though the features we use are not exactly the same as the ones in the original TRAPS publication, we still use the term TRAPS to refer to them. [6]

The input signal is passed through a bank of 1/3-octave bandpass filters, and the following processing is applied to each of the resulting subband signals. The envelope of a subband signal is calculated by squaring and low-pass filtering with 80 Hz cutoff frequency. In the lowest bands where the bandwidth is less than 80 Hz, the low-pass cutoff is lowered to match the bandwidth. The envelope signals are sampled at 400 Hz and μ-law com-

pression ($\hat{x} = \log(1 + \mu x)/\log(1 + \mu)$) is applied with $\mu = 100$. Finally, temporal difference is calculated. The motivation for applying compression and differentiation is to detect perceptually significant level changes in the subband signal. The result of this processing are the bandwise envelopes $b_i(t)$ from which the temporal features are calculated.

The actual temporal features from the envelopes are calculated by diving them into 100 ms frames with 50% overlap, and applying hamming windowing. Then a time-shift invariant representation of the envelope within each frame is desired, meaning that the position of a drum event should not have an effect on the extracted features. This is achieved by calculating the discrete Fourier transform and retaining only the magnitude spectrum. The information about the location of the event within the frame is now discarded along with the phase spectrum. The magnitude spectrum is converted to a cepstrum-like format by μ-law compressing it with $\mu = 1000$, applying discrete cosine transform (DCT), and discarding a majority of the coefficients. It was empirically noted that discarding the zeroth coefficient and retaining the following 5 produced a suitable parameterisation. The compression is used to reduce the large dynamic scale on the magnitude values before reducing the dimensionality and correlation by DCT.

Athineos et al. parameterised the envelopes by frequency-domain linear prediction in [1]. The parameterisation was efficient, but for drum transcription application is has one major drawback: the resulting features were sensitive to the absolute location of the event within the frame.

2.3 Combining the Spectral and Temporal Features

Combining the information from temporal features to the baseline HMM recogniser can be done in several ways. Hermansky et al. used a combining MLP having the outputs of the bandwise MLPs as its input to yield the final recognition result [6].

The easiest way to utilise the TRAPS features would be to concatenate the features from all bands to the feature vector used in the baseline recogniser. This approach has two major problems, however: explosion of the dimensionality of the feature vector and the highly correlated nature of the TRAPS features.

Instead of using the TRAPS features as such one could follow the example of Hermansky et al. [6] and train a detector classifier producing posterior probabilities for each target drum and for each subband. These bandwise posterior probabilities can be interpreted as features and concatenated to the HMM feature vectors. However, the HMM uses GMMs to model the features and the distribution of the posterior probabilities does not fit the model well. As a result, this approach does not produce a good result.

The solution we propose concatenates the temporal features from all bands into one feature vector and trains just one Bayesian GMM classifier for each target drum instead of an own classifier of each band. The feature vectors are subjected to PCA retaining 90% of the variance prior to training GMMs from them. In experiments it was noted that a relatively small amount of components suffices in the modelling: two components for modelling the presence of the target drum, and three for modelling the absence of the target drum. For a target drum d the GMMs produce a posterior probability $p(d)$ of the drum to be present in the frame.

As the temporal features are modelled as detectors for the target drums and the HMMs are for combinations of the drums, the posteriors for different individual drums have to be combined to one posterior for the combination. Making the assumption that the drums are independent of each other, the probability of the drum combination C can be approximated by

$$p(C) = \prod_{d \in C} p(d) \prod_{d \notin C} (1 - p(d)). \quad (1)$$

The resulting probability for this combination $p(C)$ is then added to the observation probabilities of all the states of the combination model by multiplying the probabilities before finding the optimal path through the models in decoding.

3 EVALUATIONS

The performance of the proposed system was evaluated with simulations on acoustic data. The target drums were kick drum, snare drum, and hi-hat, resulting in eight combinations to be modelled including the background model. The test material is divided into three subsets: "simple drums" consisting of simple patterns, such as 8-beat and shuffle, performed mainly with the target drums, "complex drums" containing more complex patterns and also non-target drums, and "RWC Pop" consisting of 45 pieces from RWC Popular music database [5]. The signals consisting only of drums were recorded using three different drum sets and three different recording environments. The recorded signals were processed with equalisation and multiband compression. The length of the drums-only material clips was restricted to 30 seconds, while 60 second clips were taken from the RWC songs. The setup is described in more detail in [9].

A transcribed event was judged to be correct if it deviated less than 30 ms from the event in ground truth annotations. The used performance metrics consist of precision rate, P, (ratio of correctly transcribed events to all transcribed events), recall rate, R, (ratio of correctly transcribed events to all events in ground truth), and harmonic F-measure, $F = 2RP/(P + R)$. For each material set the evaluations were run in 3-fold cross validation scheme: 2/3 of the pieces used as training material and testing with the remaining 1/3. The presented results are calculated over all folds.

To get perspective to the performance of the system, the system from [11] is used as a reference. It classifies events found by onset detector using a SVM, hence it is referred in the result tables as "SVM". [2] . The reference

[2] The used implementation was kindly provided by the MAMI consortium `http://www.ipem.ugent.be/MAMI/`.

F-measure (%)	simple drums	complex drums	RWC Pop
baseline HMM	**93.4**	84.0	66.8
HMM+TRAPS	92.9	**85.2**	**69.7**
SVM[11]	85.5	76.4	65.1

Table 1. Total average F-measures of different methods and different material sets. The presented results are calculated over all three target drums.

method was not trained for the material used in the evaluations, but instead the provided models were used. The overall results of the evaluations are given in Table 1.

Detailed results for the HMM-based systems are given in Table 2, where F-measure, precision, and recall rates are given for all three target drums for both the baseline system and the proposed extension with temporal features. It can be seen that the proposed utilisation of temporal features increases the performance slightly. Some demonstrational signals from the simulations are available at http://www.cs.tut.fi/sgn/arg/paulus/demo/.

The results for both the baseline and the reference system presented in Table 1 differ slightly from those reported in [8]. This is because some corrections were made to the ground truth annotations and longer signal excerpts were used in the evaluations.

4 CONCLUSIONS AND FUTURE WORK

We have proposed to utilise temporal features in conjunction with a HMM-based system for transcribing drums from polyphonic audio. This was shown to result in slight improvement in transcription accuracy, which is consistent with the results obtained by Hermansky et al. [6]. It was also noted that the proposed addition changed the type of the errors from insertions to deletions, which are less disturbing when listening to the synthesised transcription result.

The proposed system can be easily used as a baseline system and extended by incorporating musicological models. Such a model could be a regular N-gram model, a periodic N-gram, or a model making decision based on both the past and the future [13]. It would be preferable for the system to be able to adapt to the target signals instead of using fixed models. This could be accomplished by using the models created with the proposed method as initial models and adapting them based in the input signal.

5 REFERENCES

[1] M. Athineos and D. P. W. Ellis. Frequency-domain linear prediction for temporal features. In ASRU, St. Thomas, U.S. Virgin Islands, USA, 2003.

[2] C. Dittmar and C. Uhle. Further steps towards drum transcription of polyphonic music. In 116th AES Convention, Berlin, Germany, 2004.

material	metric	kick drum	snare drum	hi-hat
simple drums	P(%)	99.2 (92.4)	99.7 (98.2)	96.1 (94.9)
	R(%)	93.7 (98.2)	89.2 (89.4)	87.5 (90.9)
	F(%)	96.4 (95.2)	94.2 (93.6)	91.6 (92.8)
complex drums	P(%)	94.4 (87.4)	86.5 (78.5)	85.9 (82.8)
	R(%)	97.3 (97.6)	76.5 (81.1)	74.4 (78.2)
	F(%)	95.8 (92.2)	81.2 (79.8)	79.8 (80.4)
RWC Pop	P(%)	82.8 (73.1)	70.3 (52.7)	78.6 (74.1)
	R(%)	76.5 (78.7)	61.3 (66.6)	56.8 (58.6)
	F(%)	79.5 (76.8)	65.5 (58.8)	66.0 (65.4)

Table 2. Detailed results for the HMM methods. Baseline results are given in parentheses.

[3] D. FitzGerald and J. Paulus. Unpitched percussion transcription. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*. Springer, 2006.

[4] O. Gillet and G. Richard. Automatic transcription of drum sequences using audiovisual features. In ICASSP, Philadelphia, PA, USA, 2005.

[5] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In ISMIR, Paris, France, 2002.

[6] H. Hermansky and S. Sharma. TRAPS - classifiers of temporal patterns. In ICSLP, Sydney, Australia, 1998.

[7] M. F. McKinney and J. Breebaart. Features for audio and music classification. In ISMIR, Baltimore, Maryland, USA, 2003.

[8] J. Paulus. Acoustic modelling of drum sounds with hidden Markov models for music transcription. In ICASSP, Toulouse, France, 2006.

[9] J. Paulus and T. Virtanen. Drum transcription with non-negative spectrogram factorisation. In EUSIPCO, Antalya, Turkey, 2005.

[10] G. Peeters, A. La Burthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In ISMIR, Paris, France, 2002.

[11] K. Tanghe, S. Dengroeve, and B. De Baets. An algorithm for detecting and labeling drum events in polyphonic music. In MIREX, London, UK, 2005. extended abstract.

[12] B. Whitman and D. P. W. Ellis. Automatic record reviews. In ISMIR, Barcelona, Spain, 2004.

[13] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. An error correction framework based on drum pattern periodicity for improving drum sound detection. In ICASSP, Toulouse, France, 2006.

[14] S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Tech Report CUED/F-INFENG/TR38, Cambridge, UK, 1989.