DRUM TRANSCRIPTION FROM POLYPHONIC MUSIC WITH INSTRUMENT-WISE HIDDEN MARKOV MODELS

Jouni Paulus
Institute of Signal Processing
Tampere University of Technology
Korkeakoulunkatu 1, FI-33720 Tampere, Finland
jouni.paulus@tut.fi

ABSTRACT

This paper describes a system for automatic transcription of drum instruments from polyphonic music signals. For each target drum instrument, a hidden Markov model (HMM) is created to describe the sound characteristics when the instrument is played. Also, a background model with only one state is created for each instrument to describe the sound when the target instrument is not played. The signal is divided into short (2048 samples), overlapping (75%) frames and a set of features is extracted from each frame. The most likely model sequence of sound presence and absence is determined by decoding the instrument-wise HMMs with token passing algorithm.

Keywords: HMM, token passing

1 SIGNAL ANALYSIS FRONT-END

The signal analysis front-end is resembles the ones in traditional continuous speech recognition systems. The input signal is divided into frames with the length of 2048 samples (46.4 ms when the sampling rate is 44100 Hz) and 75% overlap between consecutive frames. Each frame is windowed with the Hanning function. From each frame, the following features are extracted:

- Thirteen first Mel-frequency cepstral coefficients (MFCCs), including the zeroth coefficient.
- Temporal differences of the thirteen first MFCCs, i.e., the difference from the previous frame.
- Spectral features from the first four moments, including spectral centroid, spectral spread, spectral skewness and spectral kurtosis [3]. These are calculated using a logarithmic frequency scale, as proposed in [1].
- Spectral slope and spectral roll-off, describing the general form of the spectrum.
- RMS energy in the frame.

As it is quite probable that some of the features correlate, the normalised training features are whitened with principal component analysis (PCA), and the feature vector dimensionality is reduced from 33 to 25. The number of reduced dimensions (25) was empirically chosen with few validation experiments.

2 SOUND EVENT MODELLING

For each target drum instrument, two separate HMMs [4] are constructed. The first models features in six frames of signal on the location where the instrument is hit, and the second models all the other parts of the signal. The number of frames was set to this value based on validation experiments.

The sound model consists of three states which are connected in a left-to-right manner, allowing only self-transitions and transitions to all the following states. The feature distributions in each state are modelled with Gaussian mixture models (GMMs), which are trained with EM-algorithm. The second model, which acts as a background model, has only one state and can be thought as a simple GMM. Its main purpose is to provide a likelihood value for all the observations regardless of the presence of the target drum.

The system training is done with data similar to the actual target signals, in this case with polyphonic music signals. For each drum instrument to be transcribed, all its occurrences are used to to train the "hit" model.

In transcription, the state likelihoods are evaluated based on the observations. Then the most probable sequence of the two models (instrument is present, instrument is absent) is determined with token passing algorithm [5]. The transitions between models are forced in such a manner that the same model can not be entered directly without passing through the other model. The decoding is repeated for all the target instruments and the information about their presence is combined to yield the final transcription result.

3 EVALUATION

The proposed system was evaluated with simulations in the MIREX'05 Audio Drum Detection contest. The target instruments in this case were kick drum, snare drum, and hi-hats. The system was trained with the 30-second excerpts of 23 musical pieces provided by the contest organisers.

The evaluation results can be seen in the Table 1. Each of the three test collections are on separate columns with the following abbreviations: *CD* denotes the Christian Dittmar collection, *KT* the Koen Tanghe collection, *MG* the Masataka Goto collection, and *Overall* denotes the evaluation metrics calculated over all the three col-

	CD	KT	MG	Overall
Tot Avg F	0.440	0.425	0.597	0.499
Onset P	0.558	0.558	0.629	0.596
Onset R	0.544	0.551	0.755	0.649
Onset F	0.551	0.555	0.686	0.621
BD F	0.430	0.444	0.648	0.527
HH F	0.497	0.489	0.695	0.587
SD F	0.424	0.412	0.449	0.430

Table 1: The evaluation results. See the text for explanation.

lections. *Tot Avg F* denotes the total average classification F-measure, the measure that was used to rank the systems. *Onset P* is the total overall onset precision rate, the ratio of the correctly transcribed sound onsets to the all transcribed ones. *Onset R* denotes the total overall onset recall rate, the ratio of correctly transcribed sound onsets to the ground truth, and *Onset F* is the total overall onset F-measure calculated from the two previous measures. The last three rows in the table, *BD F*, *HH F*, and *SD F* contain the F-measures calculated for each of the target instruments kick drum, hi-hats and snare drum, correspondingly.

4 DISCUSSION

The evaluation results suggest that the system requires still further development. In an informal listening test, the synthesised transcription result was compared to the original input signal. The main observation was that the system is keen to add extraneous snare drum hits. Notably often these additions were caused by singing voice. This suggests that the used features may not capture enough information of the signal content, and need still investigations. Also, the musicological modelling in the system was limited to simple bigrams controlling the transitions between the two models for each instrument. Possibly longer N-grams could be used, as suggested in [2]. Also, the number of frames that are modelled in the location of a hit, should be determined based on the properties of the modelled sound event itself, instead of using a fixed context length.

References

[1] International Organization for Standardization. *ISO/IEC 15938-4:2002 Information technology – Multimedia content description interface – Part 4: Audio*. Geneva, Switzerland, 2002.

[2] Jouni K. Paulus and Anssi P. Klapuri. Conventional and periodic N-grams in the transcription of drum sequences. In *Proc. of IEEE International Conference on Multimedia and Expo*, volume 2, pages 737–740, Baltimore, Maryland, USA, July 2003.

[3] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, Ircam, Paris, France, April 2004.

[4] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–289, February 1989.

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