Audiovisual recognition of drum sequences

Alexandre Bérard, Charles Robin

January 10th, 2014

Introduction

ENST-Drums: *30 Go* of drum audio and video sequences, played by three different drummers on their own drum kit.

Four types of sequence: *hits*, *phrases*, *soli*, *accompaniment*. All sequences are annoted, with the time of each stroke and the corresponding instrument.

Possible instruments: snare drum, bass drum, cymbals (chinese ride, crash, splash, etc.), hi-hat, toms (low tom, mid tom, etc.)

Exercise

Different possible classification tasks:

- Recognize the drummer;
- Recognize the tool used to hit (stick, brush, mallet);
- The instrument that is hit (snare drum, bass drum, etc.);
- Or a higher-level category of instrument (e.g. membranes versus plates).

We could use audio features, video features, or both of them.

Our goal

We tried three classification tasks: recognize the instrument type within three categories.

- *Super-category*: membrane, plate
- *Basic-level*: bass drum, snare drum, tom, cymbal, hi-hat
- *Sub-category*: like *basic-level*, but toms are subdivised into low tom, low mid tom, etc., cymbals into splash, ride and crash cymbals.

Bibliography

O. Gillet.
Transcription des signaux percussifs. Application à l'analyse de scènes musicales audiovisuelles.
PhD thesis, Telecom ParisTech, 2007.

O. Gillet and G. Richard.
Automatic transcription of drum loops.
In *Proceedings of ICASSP*, 2004.

O. Gillet and G. Richard.
Enst-drums: an extensive audio-visual database for drum signals processing.
In *Proceedings of ISMIR*, 2006.

P. Herrera, A. Yeterian, R. Yeterian, and F. Gouyon.
Automatic classification of drum sounds: A comparison of feature selection and classification techniques.
In *Proceedings of ICMAI*, 2002.

Whole process

1. Data selection
2. Data segmentation (extraction of strokes out of the sequences)
3. Features extraction (convert audio segments into vectors)
4. Normalization of the attributes
5. Training of a classifier (on the training data)
6. Evaluation of the classifier

Data segmentation

Audio records are sequences of strokes. We must extract those strokes. The first step is to detect the beginning of each stroke. This process is called *onset detection*. We use the time defined in the annotations as an oracle.

Then, we must define a segment size. It could be either a fixed window (e.g. 200 ms) or the whole audio sequence until the next stroke.

Feature selection

Chosen features

As suggested by Gillet et al. in [2], we used the following features:

- Means of 13 MFCC coefficients (starting by c_0), using an analysis window of 50 ms and a 50% overlap.
- 4 spectral shape parameters: spectral centroïd, width, skewness and kurtosis; defined as *SpectralShapeStatistics* in Yaafe.
- Log-energy in 6 frequency bands (chosen accordingly to the frequency content of each instrument)

Classification

Herrera et al. use a k-NN, Gillet et al. prefer a SVM. We tried both of them. Evaluation protocol: for each instrument, we measure its precision and recall. We used cross-validation, with 10 folds.

3 SVM (C=2, σ=1, 10 folds)

Instrument	Precision	Recall	F-measure
Bass drum	90.6%	76.1%	82.7%
Snare drum	92.8%	82.6%	87.4%
Hi-hat	84.7%	94.1%	89.2%
Average	89.4%	84.3%	86.4%

Remarks

Good results, because we skipped the *onset detection* step by using an oracle.

There is room for improvement: feature selection, parameters tuning, etc. We didn't evaluate our system on noisy data.

Thank you for listening. Questions?