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 several classification tasks, using audio features. The goal is to recognize the instrument type within four taxonomies.

- *Super-category*: membrane, plate
- *Gillet's taxonomy*: bass drum, snare drum, hi-hat (75% coverage)
- *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.

Summary

1. Data segmentation
2. Feature extraction
3. Feature selection
4. Classification
5. Results and conclusion

Data segmentation

Audio records are sequences of strokes. We must extract those strokes.

- Detect the beginning of each stroke. This process is called *onset detection*. In our case, we use the time defined in the annotations as an oracle.
- Define a segment size. It could be either a fixed window (e.g. 200 ms) or the whole audio signal until the next stroke.
- We decided that instruments that are played within the same window of 50 ms belong to the same segment.

Feature computation

We used *Yaafe* to compute the features of each audio segment. For each of those features we compute the mean value, with an analysis window of 50 ms, and 50% overlap.

Feature list

Feature	Yaafe name	Dimension
Mel Frequency Cepstrum Coef	MFCC	13
Spectral shape parameters	SpectralShapeStatistics	4
Temporal shape parameters	TemporalShapeStatistics	4
Energy ratio in octave frequency bands	OBSIR	9
Total energy	Energy	1
Zero crossing rate	ZCR	1
Linear prediction coefficients	LPC	6
Spectral flatness	SpectralFlatness	1
Perceptual sharpness	PerceptualSharpness	1
Perceptual spread	PerceptualSpread	1

We used 3 sets of features : all features (*all*), manually chosen features: mfcc and spectral shape parameters (*manual*), and automatically selected features (*auto*).

Feature selection: IRMFSP

We used the algorithm *IRMFSP*. Get the attributes that have the maximum inter-classes distance, and minimum intra-class distance. Feature selection is a dimensionality reduction.

Table : First 4 selected attributes

Instrument	Attributes
Bass drum	$OBSIR_3$, $OBSIR_2$, $MFCC_0$, $MFCC_1$
Snare drum	$OBSIR_2$, $MFCC_2$, $SpecShape_3$, $Spread_0$
Hi-hat	LPC_0, $TempShape_2$, $MFCC_4$, $OBSIR_8$

Classification

Given a taxonomy with N instruments, determine for each of the instruments, whether it is played in the audio segment or not.

There are two approaches:

- A single classifier, with 2^N classes (one class for each combination of instruments.)
- N binary classifiers, which decide if the instrument is played or not.

Evaluation protocol

For both approaches, we tried two classifiers:

- SVM classifier (with RBF kernel) as *Gillet* proposed
- K-NN classifier, like *Herrera*, with $k = 5$

We used a 10-folds cross-validation. We reported, the *precision*, *recall* and *f-score* for each instrument.

$$precision = \frac{correct}{predicted},\ recall = \frac{correct}{true},\ f1 = \frac{2 \times P \times R}{P + R}$$

Results

Results for SVM, with *manual* features:

Table : SVM (C=2, σ=1)

Instrument	Precision	Recall	F1
Bass drum	91.4%	74.1%	81.9%
Snare drum	93.0%	79.6%	85.8%
Hi-hat	84.7%	93.2%	88.8%
Average	89.7%	82.3%	85.5%

Table : 3 SVM (C=2, σ=1)

Precision	Recall	F1
90.6%	76.1%	82.7%
92.8%	82.6%	87.4%
84.7%	94.1%	89.2%
89.4%	84.3%	86.4%

Results

Results for binary SVM and K-NN, with *all* features:

Table : 3 K-NN (K=5)

Instrument	Precision	Recall	F1
Bass drum	88.5%	84.9%	86.6%
Snare drum	91.9%	90.8%	91.3%
Hi-hat	91.6%	92.7%	92.1%
Average	90.7%	89.5%	90.0%

Table : 3 SVM (C=2, σ=1)

Precision	Recall	F1
93.4%	75.9%	83.7%
95.5%	82.3%	88.4%
84.6%	97.1%	90.4%
91.2%	85.1%	87.5%

Results

Results for binary SVM, with *all* and *auto* features:

Table : 3 SVM (*auto*)

Instrument	Precision	Recall	F1
Bass drum	86.6%	86.2%	86.4%
Snare drum	90.6%	86.3%	88.4%
Hi-hat	82.4%	94.4%	88.0%
Average	86.5%	89.0%	87.6%

Table : 3 SVM (*all*)

Precision	Recall	F1
93.4%	75.9%	83.7%
95.5%	82.3%	88.4%
84.6%	97.1%	90.4%
91.2%	85.1%	87.5%

Results

Average results for SVM, with *manual* features:

Taxonomy	Precision	Recall	F1
Gillet(N=3)	89.7%	82.3%	85.5%
Basic-level(N=5)	87.7%	65.8%	72.7%
Sub-category(N=12)	81.3%	44.4%	52.3%

Remarks

In Gillet's thesis, $F1 = 69.8\%$. Our results are comparatively good, because we skipped the *onset detection* step by using an oracle, and our evaluation protocol is less strict.

There is room for improvement:

- Better feature selection
- Parameters tuning
- Different type of classifier for each instrument
- Use video features.