MIR TASK: Beatbox Dataset Evaluation Using Automatic Drum Transcription

Pedro Ricardo
UPF Master Student
pedromanuel.ricardo01@estudiant.upf.edu

1. INTRODUCTION AND TASK PROPOSAL

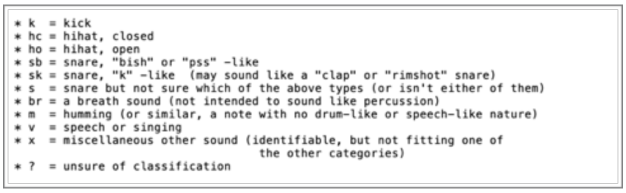
According to MIREX2018, drum transcription is defined as the task of detecting the positions in time and labeling the drum class of drum instrument onsets in polyphonic music. The article *A Review of Automatic Drum Transcription* [1] offers an in-debt explanation about automatic drum transcription as well as an extremely up to date state of the art.

The present task proposes to test and evaluate a Beatbox dataset using an automatic drum transcription algorithm - ADTLib. To do so, a second dataset ENST-Drums (used initially to develop ADTLib) will be used as a term of comparison. Further conclusions about the Beatbox dataset and its elements will be made.

2. DATA

2.1 BeatBoxset 1

The Beatboxset1 audio data set available in the Audio Content Analysis contains beatboxing recordings from various contributors, who recorded the clips themselves in various conditions. A spreadsheet file "beatboxset1.csv" accompanying the dataset provides metadata for the recordings. Further annotations of the recordings are also included: these mark the positions of onsets as well as categorizing the events into a handful of standard classes (Fig 1)



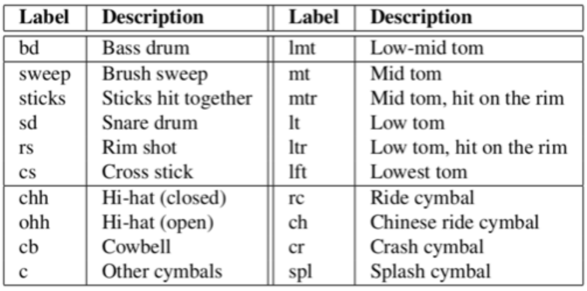
```
* k  = kick
* hc = hihat, closed
* ho = hihat, open
* sb = snare, "bish" or "pss" -like
* sk = snare, "k" -like  (may sound like a "clap" or "rimshot" snare)
* s  = snare but not sure which of the above types (or isn't either of them)
* br = a breath sound (not intended to sound like percussion)
* m  = humming (or similar, a note with no drum-like or speech-like nature)
* v  = speech or singing
* x  = miscellaneous other sound (identifiable, but not fitting one of
                                  the other categories)
* ?  = unsure of classification
```

Figure 1. Labels used for the Beatboxset1 dataset

In order to compare the handmade annotations and the output of the ADTLIB, the previous categorization will be reduce to three different categories: Kick (k), Snare (sb, sk, s) and Hihat (hs, ho)

2.2 ENST-Drums

The ENST-Drums database is a large and varied research database for automatic drum transcription and processing. The database is composed of different multichannel recordings from three drummers on three different drum kits. For each drummer, the data set provides individual hits and phrases, individual soli which are more complex than the phrases and longer tracks played without scores but with an accompaniment.



Label	Description	Label	Description
bd	Bass drum	lmt	Low-mid tom
sweep	Brush sweep	mt	Mid tom
sticks	Sticks hit together	mtr	Mid tom, hit on the rim
sd	Snare drum	lt	Low tom
rs	Rim shot	ltr	Low tom, hit on the rim
cs	Cross stick	lft	Lowest tom
chh	Hi-hat (closed)	rc	Ride cymbal
ohh	Hi-hat (open)	ch	Chinese ride cymbal
cb	Cowbell	cr	Crash cymbal
c	Other cymbals	spl	Splash cymbal

Figure 2. Labels used for the ENST-Drums dataset

In order to match the amount of data available in Beatboxset1, a subset of ENTS- Drums were chosen. Fourteen tracks across different genres and tempos was chosen. A complete list of the tracks is available with the Jupyter Notebook were the task was implemented.

3. ADTLIB

Automatic Drum Transcription Library (ADTLib) is a library that contains open source ADT algorithms to aid other researchers in areas of music information retrieval (MIR). The algorithms returns both a .txt file of kick drum, snare drum, and hi-hat onsets location.
ADTLib was develop as an open source implementation of the different systems and algorithms initially presented in the research paper *Automatic drum transcription for polyphonic recordings using soft attention mechanisms and convolutional neural networks* (2017) [2]. The paper by Carl Southall, Ryan Stables and Jason Hockman focus in implementing different techniques using soft attention mechanisms (SA), convolutional neural networks (CNN) and bidirectional recurrent neural networks (BRNN) on the automatic drum transcription problem. However, the available implementation of ADTLib does not train neural networks. The music files are processed through a pre-trained neural network to give the automatic drum transcriptions. The scheme presented in Figure 3 tries to give an overview about the implementation of the ADTLib library. More information can be found in the original paper and in the blog post *Notes on Tensorflow and how it was used in ADTLib.*

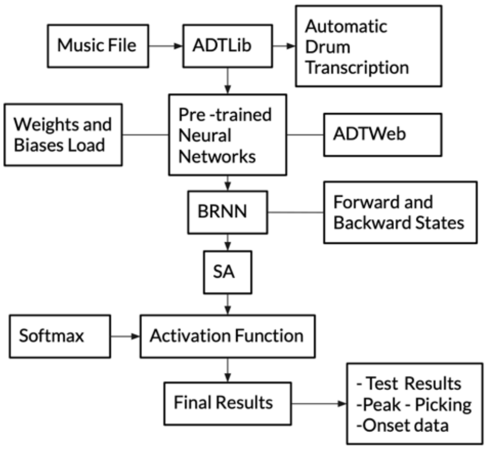


Figure 3. Labels used for the Beatboxset1 dataset

4. EVALUATION METHOD

The evaluation method focus in understanding how well the implemented algorithm performs by comparing the hand-made annotations, available in the data sets, to the annotations produced by ADTLib. The evaluation methodology it's the same used in *Automatic drum transcription for polyphonic recordings using soft attention mechanisms and convolutional neural networks.* A F-measure value will be computed for each different sound: kick, snare and hi-hat. The detected onsets are accepted as true positives if they fall within 50 milliseconds of the hand-made annotations provided by each dataset. The implementation of the F-measure was done using the library mir_eval. More information about mir_eval can be red in mir_eval: *A Transparent Implementation of Common MIR Metrics* [3]

5. RESULTS

Four different graphs were computed: F-measure score across Beatbox dataset (Fig. 4) and F-measure score across ENTS dataset (Fig. 5), F-measure across different instruments (Fig. 6) and F-measure across datasets (Fig. 7).

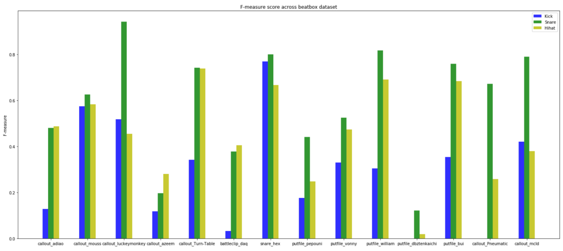


Figure 4. F-measure score across Beatbox dataset

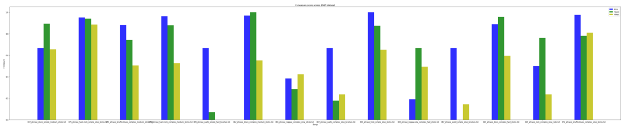


Figure 5. F-measure score across ENTS dataset

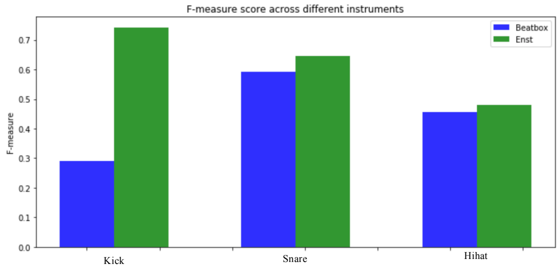


Figure 6. F-measure across different instruments

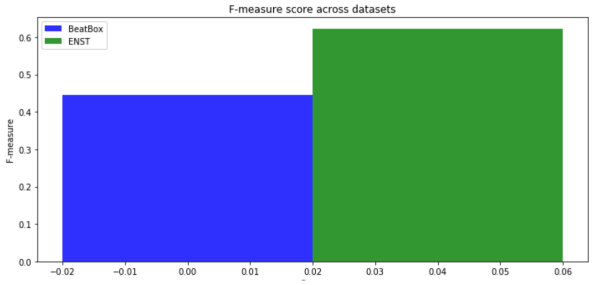


Figure 7. F-measure across datasets

Both Figure 4 and Figure 5 show the F-measures across each instrument (Kick, Snare Hihat), across the different tracks that are included in the different datasets. Across both datasets, there are tracks were at least one of the instruments doesn't have a value. Listening tests show that that in the case of the Beatbox dataset the two tracks that don't show any kick information (putfile_dbztenkaichi and callout_Pneumatic) have substantial harmonic information.

Figure 7 shows the mean of the F-measure of each dataset. The calculation of this value can be seen in the Jupyter Notebook. Because the ENTS dataset was used to develop the ADTLib its was expected to achieve higher results.

The most interesting results come from analyzing Figure 6. Both the snare and the hihat have very similar values across the two datasets. It is safe to say that between all instruments the snare and the hihat sounds made by the beatboxers are similar the snare and hihat sounds of the real drummer. The same conclusion can't be when it comes to the kick.

CONCLUSIONS AND FUTURE WORK

Although the data used for this experience was quite reduced, the ENTS-drum dataset performed better in comparison to the Beatbox1 dataset. The snare and the hihat achieve identical levels of similarity between both datasets. Although a final conclusion about the reproducibility of drum sounds by the human voice can't be made only having in consideration the results of this project, further study seems worthy.

For future work it would be interesting to do mappings across different genres and tempos.

Further notes, links to each reference, dataset and libraries used in this project can be found in the Jupyter Notebook

6. REFERENCES

[1] C. Dittmar, C. Southall: "A review of automatic drums transcription" **in** IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 26 , Issue: 9, Sept. 2018)
[2] C. Southall, R. Stables, J. Hockman: "Automatic Drum Transcprition For Polyphonic Recordings Using Soft Attention Mechanisms And Convolutional Neural Networks" in ISMIR, 2017
[3] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis, "Mir eval: A transparent implementation of common mir metrics," in ISMIR, 2014.