BEATBOX CLASSIFICATION USING ACE

Elliot Sinyor **Cory McKay** **Rebecca Fiebrink** **Daniel McEnnis** **Ichiro Fujinaga**
Music Technology
McGill University
Montreal, Quebec

{elliot.sinyor, cory.mckay, rebecca.fiebrink, daniel.mcennis}@mail.mcgill.ca
ich@music.mcgill.ca

ABSTRACT

This paper describes the use of the Autonomous Classification Engine (ACE) to classify beatboxing (vocal percussion) sounds. A set of unvoiced percussion sounds belonging to five classes (bass drum, open hihat, closed hihat and two types of snare drum) were recorded and manually segmented. ACE was used to compare various classification techniques, both with and without feature selection. The best result was 95.55% accuracy using AdaBoost with C4.5 decision tress.

Keywords: ACE, beatboxing, classification, feature selection.

1 INTRODUCTION

Besides tapping one's fingers, vocalizing percussion is perhaps the most intuitive way for musicians and non-musicians alike to express a rhythm. The range of sounds that can be made by one's mouth, however, is far greater than that of fingers alone. The act of vocalizing percussive sounds is as old as music itself, and almost every culture has its own approach. A notable example is Indian Tabla players' use of *bols*, a set of vocal sounds used to express rhythmic phrases.

In North American culture, two examples immediately come to mind: 50's doo-wop, and more recently, beatboxing. Both originated in African-American music. The term "beatboxing" originally referred to the mimicking of early 80's drum-machines, also known as beatboxes. While beatboxing was first used as a backing rhythm for rap performance, it has been developed into an art form in and of itself by performers like Biz Markie and Rahzel.

In MIR research, the main interest in beatboxing has been in using it as a means of querying stored drum data, but other possibilities exist. Reliable recognition of different drum sounds could serve as a starting point for developing an intuitive rhythm-performance interface. Similarly, it could be used as an input to a metrical analysis system. The methods described here can also be used to develop a more general mouth-based control channel, as the sounds require very little effort to produce and do not necessarily need to be mapped to drum sounds, or even sounds for that matter.

This project centres around an attempt to reliably classify five different drum sounds: bass or kick drum, closed hihat, open hihat and two types of snare drum. We began by collecting a set of vocal percussion samples, from both expert beatboxers as well as non-beatboxers. The recording of beatboxers was carried out primarily as a study on common ways to vocally express drum sounds.

For classification, we ran our collected data through the Autonomous Classification Engine (ACE) described in [1]. ACE combines several approaches to classification and can be used to determine which classifiers and features are effective at classifying a given data set. Furthermore, we used a k-nearest neighbour classifier coupled with genetic algorithm (GA) based feature selection (described in [2]) as a baseline to compare with ACE's performance.

2 RELATED WORK

There are numerous publications dealing with classification of instrument sounds and several dealing specifically with drum sounds (e.g., [3]). This problem also closely resembles speech-recognition problems. Since the audio signals in question are for the most part unvoiced and extremely short (20–100ms), pitch-based analysis tools are generally not successful. This rules out many phoneme-based techniques. Approaches based on plosives and fricatives [4–6], however, are relevant. Generally, previous attempts have used a number of timbral features and statistical classifiers.

In [7], the authors describe a system to retrieve a MIDI drum loop from a bank of recorded drum loops by means of a user beatboxing into a microphone. This system attempts to classify the drum sounds into one of three categories, namely bass drum, snare, and hihat. A 97.3% accuracy is reported using zero-crossing rate as the sole feature. While this result is impressive, simply using zero-crossing rate is unlikely to yield similarly high results for more classes or for a larger number of subjects.

In [8], the author describes a variety of spectral and temporal features to classify beatboxing samples into four classes: bass drum, snare drum, closed hihat and open hihat. A success rate of 86% for the training set is

achieved using 24 features and a C4.5 classifier with boosting. Interestingly, the author reports 90% using the same classifier for a previously unseen test set.
Somewhat different, but worth mentioning, is the system described in [9], which uses complete syllables (e.g., "don", "ta", "zur") to represent drum sounds. Each syllable is subdivided into consonants, vowels, and nasal sounds.

3 DATA COLLECTION

3.1 Recording

The data collection involved three accomplished[1] beatboxers and three non-beatboxers. The rationale behind using beatboxers was to observe the range of sounds they made and how they opted to mimic the four initial classes we required. While we could have done the same with non-beatboxers, they are less likely to have found and refined a particular way to make a certain drum sound. These preliminary observations proved useful, as we discovered two common ways to make a snare sound, which led us to change our taxonomy.

Each subject recorded individual drum hit sounds both separately and as part of beat patterns. Subjects were instructed to imitate a kick drum (also known as a bass drum), a snare drum, a closed hihat and an open hihat. They were provided with an example of each but told to vocalize as was usual for them. Also, they were instructed to make the sounds unvoiced, or non-pitched.

The recordings took place in two slightly different acoustic environments. Half were done in a small office with linoleum flooring and painted concrete walls, and the remainder were done in an office-like environment with carpeting and acoustic-tiled ceilings. The rooms likely had only a small effect on the recordings, since almost all subjects held the microphone extremely close to their mouths, often touching it to their lips. In fact, two of the beatboxers made use of their free hand to cup the microphone while doing bass drum hits.

The recording was done using ProTools with a Digidesign MBox audio interface and a Shure SM58 dynamic vocal microphone. The microphone input levels were kept the same for all recordings, but no normalizing was done on any of the audio data, as some variance in level was desired to account for the difference in beatboxing style from subject to subject.

The recording was split into two parts. First, subjects were told to simply make beats as they pleased using the above drum sounds. Next they were told to record each drum hit in sets of 10, and then repeat this process three more times, yielding 40 samples of each type. They were instructed to try to keep the hits similar to each other.

3.2 Segmentation

The segmentation was done manually using Audacity, an open-source and multi-platform audio editor. Special effort was made to include little to no silence, as the resulting low-amplitude background noise could bias certain features, most notably ones based on the zero-crossing rate. Once segmented and labelled, all drum hits were exported as numbered WAV files, all mono with a bit rate of 16 bits and a sampling rate of 44.1 kHz.

While manual hit segmentation was sufficient for the needs of this project, larger-scale projects would require automatic segmentation. Several experiments on our data showed that the relative difference function, as discussed in [10], is particularly useful for onset detection.

The entire sample set consisted of 1206 drum hits. After auditioning the sample set, 12 samples were removed due to audible clicks or very low level input. Since the goal was to have a realistic data set, some questionable samples were included and only obviously flawed ones were removed.

In general, each drum hit sounded qualitatively similar across subjects. This may have been partly due to the fact that the subjects were given model examples. The kick drums most closely resembled the unvoiced bilabial plosive (/p/), the closed hihat resembled the unvoiced alveolar plosive (/t/) and the open hihat most resembled the unvoiced alveolar fricative (/s/).

The snare hits were vocalized slightly differently by different subjects. Two of the subjects imitated the snare drum by combining the bilabial plosive and alveolar fricative to make a short and explosive "pss" sound. It can be thought of as a combination of the /p/ sound with the /s/ sound, as the waveform in Figure 4 shows. The remaining beatboxer imitated the snare drum by making an unvoiced velar plosive, or /k/ sound. Two of the non-beatboxers also did this, and the remaining one was unsure of what to do, so he did both. The snare category was thus subdivided into two: p-snare and k-snare. The figures below provide illustrations of the waveforms from the five classes.

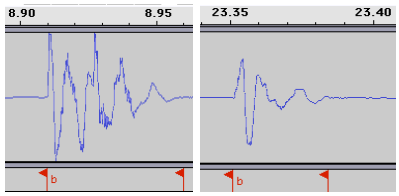

Figure 1. Kick drum (beatboxer, non-beatboxer)

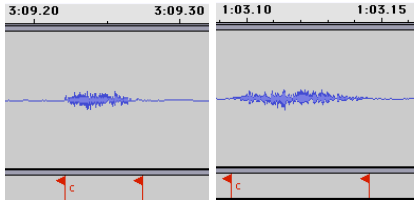

Figure 2. Closed-hihat (beatboxer, non-beatboxer)

[1] All three perform as part of *a cappella* groups and two participate in beatboxing competitions.

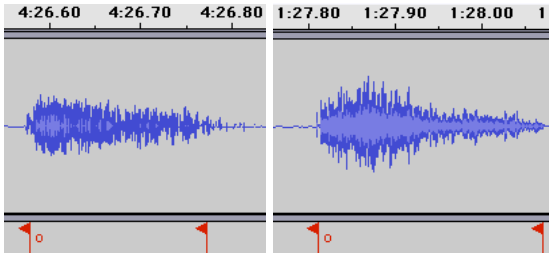

Figure 3. Open-hihat (beatboxer, non-beatboxer)

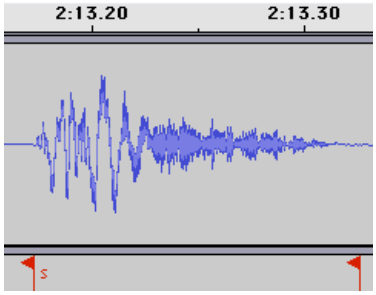

Figure 4. p-snare (beatboxer)

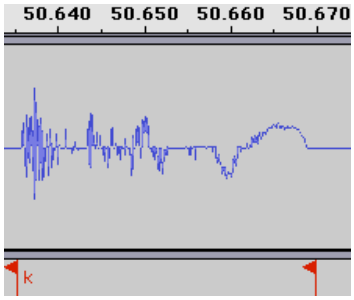

Figure 5. k-snare (non-beatboxer)

4 CLASSIFICATION

The final dataset, totalling 1192 samples, is divided into five classes: 311 kick drum, 298 closed hihat, 290 open hihat, 137 p-snare and 156 k-snare.

4.1 Features

The following features were extracted using ACE's jAudio feature extractor component [11]:

```
Sp_Centroid_Overall_Avg                    false
Sp_Centroid_Overall_Std_Dev
      false
Sp_Rolloff_Point_Overall_Avg
      true
Spectral_Rolloff_Point_Ovl_Std_Dev         true
Sp_Flux_Overall_Avg                        false
Spectral_Flux_Overall_Std_Dev
      false
Compactness_Overall_Avg                    true
Compactness_Overall_Std_Dev                true
Spectral_Variability_Overall_Avg           true
Spectral_Variability_Overall_Std_Dev       true
RMS_Overall_Avg                            false
RMS_Overall_Std_Dev                        true
RMS_Derivative_Overall_Avg                 true
RMS_Derivative_Overall_Std_Dev             true
ZC_Overall_Avg                             true
ZC_Overall_Std_Dev                         true
ZC_Derivative_Overall_Avg                  true
ZC_Derivative_Overall_Std_Dev
      false
Strongest_Freq_Via_ZC_Overall_Avg          false
Strongest_Freq_Via_ZC_Overall_Std_Dev      false
Strongest_Freq_Via_SC_Overall_Avg
      true
Strongest_Freq_Via_SC_Overall_Std_Dev      false
Strongest_Freq_Via_FFT_Max_Overall_Avg     false
Strongest_Freq_Via_FFT_Max_Ov_Std_Dev      true
```

Figure 6. Features used for classification.

In the above list, RMS refers to Root Mean Sqaure, ZC refers to zero-crossing, and SC refers to spectral centroid. `Strongest_Freq_Via_ZC` refers to the strongest frequency in Hz that corresponds to the ZC rate. Likewise, `Strongest_Freq_Via_FFT_Maxds` refers to the frequency corresponding to the highest peak of the FFT. Flux is a measure of the difference between two successive FFT windows. Compactness is used to measure the degree of noise in a signal and is measured as follows:

$$\sum_{n=1}^{N-1} \log(M[n]) - \frac{\log(M[n-1]) + \log(M[n]) + \log(M[n+1])}{3}$$

where M[n] is the *n*th bin of the magnitude spectrum. For FFT-based features, a Hanning window of 512 samples with no overlap was used.

4.2 Feature Selection

In addition to the ACE-based experiment, a second experiment was performed using a genetic algorithm feature-selection system described in [2]. The chosen features are labelled as "true" in Figure 6. These features were chosen using an initial population of 50 chromosomes, and it took 14 generations to achieve convergence. Since the GA assigns fitness to chromosomes based on the performance of a classifier that is trained with the corresponding feature set, the chosen features can be quite classifier-dependent. In our case, a k-NN classifier was used, for which the presence of redundant features can impair classification. Removing redundancies has a positive effect on classification rate and thus chromosomes that exclude redundant features will have higher fitness scores. What this means is that the selected features are more likely to be discriminant, however the rejected features are not necessarily useless.

4.3 Results using ACE

Without feature selection, the best accuracy rate achieved was 95.55% using AdaBoost with C4.5 decision trees as base learners. Other successful approaches used by ACE included a backpropagation neural network, which yielded

[2] All three perform as part of *a cappella* groups and two participate in beatboxing competitions.

an accuracy rate of 93.37%, and a support vector machine, which yielded a rate of 83.8%. ACE also performed naïve Bayes and k-NN classification, with k ranging from 1 to 12. Also, as a means of comparing our system with the one described in [1], we reduced the number of classes to three (bass, snare, hihat) and obtained an accuracy rate of 98.15%.

Using the features selected by the GA system with a 1-NN classifier, an accuracy of 94.55% was achieved. This can be compared to a rate of 89.36% when ACE used a 1-NN classifier and all 24 features. In all cases, 10-fold cross validation was used.

Table 1. Confusion matrix for best classification result, AdaBoost with C4.5

Classified as:

a	b	c	d	e	Actual:
309	0	1	0	1	a = kick
0	278	12	0	0	b = open
0	15	273	6	4	c = closed
0	1	5	149	1	d = k-snare
2	1	1	3	130	e = p-snare

Table 2. Recent attempts at beatbox classification. Only best accuracy rates are shown.

Author	# of classes	# of samples	# of feat.	Classifier	Acc.
Kapur	3	75	1	ANN	97.3%
Hazan	4	242	28	C4.5 w/ boosting	86%
Sinyor	5	1192	24	C4.5 w/ boosting	95.55%
Sinyor	3	1192	24	C4.5 w/ boosting	98.15%

5 CONCLUSION AND FUTURE WORK

The results described above are promising, and the high classification rate shows this work to be a good starting point for future vocalized music query systems and other voice-control applications. The GA-based feature selection approach identified particular features that would serve as a focus for future work with this type of data. It should be noted that some sort of automatic segmentation would be required in order for this approach to be used as part of a larger system.

6 ACKNOWLEDGEMENTS

The authors would like to thank beatboxers Benjamin Hammond, Jason Levine, and Kweku Sam Kwofie as well as non-beatboxers Ansel Brandt and Joseph Malloch for their time and beats.

REFERENCES

[1] McKay, C., Fiebrink, R., McEnnis, D., Li, B., and Fujinaga, I. "ACE: A framework for optimizing music classification," *International Conference on Music Information Retrieval*, 2005.

[2] Fiebrink, R., McKay, C., and Fujinaga, I., "Combining D2K and JGAP for efficient feature weighting for classification tasks in music information retrieval," *International Conference on Music Information Retrieval*, 2005.

[3] Tindale, A., Kapur, A., Tzanetakis, G., and Fujinaga, I. "Retrieval of percussion gestures using timbre classification techniques," *Proceedings of the International Conference on Music Information Retrieval*, 2004.

[4] Chan, C., and Ng, K. "Separation of fricatives from aspirated plosives by means of temporal spectral variation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(15) (Oct. 1985), 1130–1137.

[5] Molho, L. "Automatic acoustic-phonetic analysis of fricatives and plosives," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1976.

[6] Demichelis, P., De Mori, R., Laface, P., and O'Kane, M. "Computer recognition of stop consonants," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1979.

[7] Kapur, A., Benning, M., and Tzanetakis, G. "Query- by-beat-boxing: Music retrieval for the DJ," *Proceedings of the International Conference on Music Information Retrieval*, 2004.

[8] Hazan, A. "Towards automatic transcription of expressive oral percussive performances," *Proceedings of the International Conference on Intelligent User Interfaces*, 2005.

[9] Nakano, T., Ogata, J., Goto, M., and Hiraga, Y. "A drum pattern retrieval method by voice percussion," *Proceedings of the International Conference on Music Information Retrieval*, 2004.

[10] Klapuri, A. "Sound onset detection by applying psychoacoustic knowledge," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[11] McEnnis, D., McKay, C., Fujinaga, I., and Depalle, P. "JAudio: A feature extraction library," *International Conference on Music Information Retrieval*, 2005.