[[1]](#footnote-1)

Automatic Recognition of Beatboxing Sounds using a Hidden Markov Model

Ankoor Apte and Sidney Cozier

*Abstract*— Beatboxing or vocal percussion is a form of music that, despite its growing relevance in popular culture, has not been studied extensively in the Music Information Retrieval (MIR) community. This paper proposes a beatboxing recognition system for automatic music transcription. The machine-listening system consists of a Hidden Markov Model (HMM) that is used to predict the types of sound present in a beatboxing query recording. Additionally, onset detection is used to determine the locations of each of the sounds in the query, providing a basis for transcription. The HMM is constrained to having 10 states, i.e. it is capable of recognizing a pre-defined set of 10 percussion sounds. The system is implemented in Python and tested on a dataset of 60 recordings (over 1000 sounds). Results. Future work includes acquiring more data, processing the database further and developing a more robust feature vector.

# Introduction

## A. Motivation

Beatboxing can be defined as the art of vocal percussion, or mimicking real drum machines with the human mouth and voice. Beatboxing performances are done solo or with a group of singers, like in a cappella music. Due to an inherent dependence on physical features (how the human mouth is shaped) and subjective interpretation (how a drum sound is translated to a vocal sound by each beatboxer), beatboxing sounds tend to vary significantly across individual performers. Despite this lack of consistency, most beatboxing sounds across performers can be loosely recognized as standard drum sounds (e.g. kicks, snares, hi-hats, among others). Therefore, it is possible to characterize or recognize beatboxing sounds as standard drum sounds. The solution in this paper attempts to automate this process.

## B. Problem Statement

Given an audio recording of a solo beatboxing performance, design a system that recognizes the percussion sounds (using a pre-learned set of 10 percussion sounds) and identifies their onset locations in the recording.

# System Design

The overall system requirements are that it must take in an input audio recording (of beatboxing) as a query, and produce an output list that includes the label and time of onset for every percussion sound in the recording. In this design, the output is formatted as a list of tuples, where each tuple specifies the sound and onset time.

## A. Hidden Markov Model

The proposed approach for beatboxing sound recognition is a Hidden Markov Model (HMM). The overall system is shown below.

An HMM was chosen as the final design due to the relative simplicity of the implementation, and the ability to condense the test data into feature vectors. The HMM design includes observations, states and a model. In this case, the observations are sections of the input audio signal (or query) that capture each individual sound. There are 10 distinct states that correspond to the 10 pre-defined set of percussion sounds. Finally, the model is a multivariate normal probability distribution that can be used to estimate the hidden state.

## B. Percussion Dictionary

The HMM has 10 distinct states that correspond to sounds in the percussion dictionary. The percussion dictionary was defined from a set of typical Western drum sounds, using Standard Beatboxing Notation (SBN).

1. Percussion Dictionary

| Percussion Sound | Standard Beatboxing Notation |
| --- | --- |
| Kick | B |
| Snare 1 | Pf |
| Snare 2 | Pch |
| Snare 3 | K |
| Snare 4 | ^Ksht |
| Snare roll | rrh |
| Closed hi-hat | t |
| Open hi-hat/crash/cymbal | ts |
| Rimshot | k |
| Lip Oscillation | BB |

## C. Obtaining Observations

Given the input audio query, the system obtains an array of observations and calculates feature vectors for each observation. The goal of this step is to isolate each sound (in the samples domain) made by the beatboxer and summarize its content in a feature vector. First, the onset locations in the query are calculated. Then, the tempo of the query is estimated. The tempo is used to calculate the window size of the observation, i.e. the number of samples after the onset that are recorded in the observation. The window size is calculated to be one eighth measure long. Figure\_\_ shows an example of an observation, obtained from a longer recording of a beatboxing recording.

After obtaining an array of such observations from the input query, the system calculates feature vectors for each observation. The feature vector contains 8 parameters that describe the observation sound: attack, release, sustain, decay, mean amplitude, maximum intensity value on a spectrogram of the sound, and the frame and frequency index of the maximum. These 8 parameters were chosen after initial data exploration – it was clear that the ADSR envelope would be an effective way of characterizing each percussion sound. Figures \_\_\_ show examples of ADSR envelopes for a kick, snare…

## D. Training Model

The model is trained from recordings of 6 different beatboxers making each of the 10 percussion sounds 4 times (i.e. a total of 240 sound observations). Features vectors are calculated for each of the 240 sounds. By combining the known ground truth with the feature vectors, mean and covariance parameters can be calculated for each percussion sound. Therefore, a multivariate normal distribution is obtained for each state, which is then used as the model for the HMM.

## E. Estimating States

Given a model and an array of observations, the system constructs a pairwise similarity matrix of size (10, N) where N is the number of observations in the array. The pairwise similarity matrix value at index (i, j) specifies the log of the conditional pdf value of observing the j-th feature vector given that the percussion sound was of type i. Once this matrix is constructed, the maximum probability state is picked in each column as the estimated hidden state.

## F. Key Assumptions

This system approaches the problem statement using a few assumptions that are worth noting:

* There is only one sound per eighth measure of a beatboxing recording.
* The beatboxing recording has constant tempo.
* The transition matrix for the HMM is constant, to avoid biases in the sound

# Data Collection

## A. Recording Procedure

The recording procedure involves 10 recordings, as described below:

* One recording at 120 BPM with each sound in the percussion dictionary performed 4 times (for model training)
* Three recordings of beat pattern A at 80/100/140 BPM
* Three recordings of beat pattern B at 80/100/140 BPM
* Three recordings of pseudo-random beats (the beatboxer chooses from a set of three) or random beats (the beatboxer makes up their own beat)

The complete beat patterns are available in the Appendix. The recording procedure was designed to get a spread of data (known and unknown tempo, known and unknown beat patterns). The first recording was specifically designed for model training.

## B. Data Collection Status

The current database (used for analysis in this paper) includes recordings from 6 different beatboxers, i.e. 60 recordings that collectively consist of over 1000 sounds. With the recording procedure available online, we expect that people will continue to contribute to this database.

# Results

After obtaining the data, we tested each of the recordings according to the ground truth for that recording. In order to capture a broad picture of how the system performs, we created three separate types of ground truths. The majority of the recordings (6 per person) have "absolute" ground truths: we know the exact beat pattern and tempo in the recording. The other two recordings have semi-absolute ground truths: one where tempo is unknown, and one where tempo is unknown and the beat pattern is one of three possibilities. By using the semi-absolute ground truths, we created a range of variations on the fundamental problem of automatic transcription.

We then analyzed each recording based on its ground truth, which resulted in three separate categories of results: absolute ground truth results, free tempo results, and multiple choice results. The absolute ground truth results show how well the system transcribes the order of sounds when the recording’s tempo was fixed. The free tempo results are similar to the absolute ground truth results; however, the ground truth is constructed based on the computer-estimated tempo, rather than a known tempo. Both of these types of results are given as match rates, which are the percentage of sounds predicted correctly for a particular recording. The multiple choice results are slightly different. We calculated match rates on the three possible beat patterns that the beatboxer might have performed and used the maximum match rate as a prediction of which pattern was actually performed. Thus, the results for the recordings with multiple choice ground truths are predictions of entire recordings, not individual sounds.

The performance of the system decreased drastically for the recordings with higher tempo and/or with less certain ground truth information, which matched our expectations. For example, one of the more basic beat patterns had a mean match rate 56.7% for a tempo of 80 bpm (recording 2). However, when the tempo was increased to 140 bpm, the mean match rate dropped to around 28%. The increase in tempos dropped the system performance presumably. Additionally, the performance also suffered for the recordings with unknown tempo. However, our results are much better for the aided transcription.

# Discussion

Decreasing match rate with increasing tempo

Worse results for unknown tempo

## A. Plots (if we need them for further analysis

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation.

## B. Limitations

Headings, or heads, are organizational devices that guide the reader through your paper.

## C. Future work

Get more data

Expand to include non-constant tempo

Better feature parameters

Solidify observations

# Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Appendix

Beat patterns:

Acknowledgment

We would like to thank Professor TJ Tsai for his guidance throughout the semester and during the project. We also acknowledge the time and effort put in by the six beatboxers from the Claremont Colleges a cappella groups to record the test data.

References

1. G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems* (Book style)*.* Belmont, CA: Wadsworth, 1993, pp. 123–135.

1. [↑](#footnote-ref-1)