Beatboxing to Drums using Support Vector Machines

Arvind Ramanathan
University of Rochester
Electrical and Computer Engineering
arm@ur.rochester.edu

ABSTRACT

This paper elucidates a method to automatically covert a beatboxing audio into an MIDI synthesized audio waveform. The method is to detect the onsets of a recorded beatboxing audio and segment the audio into individual slices and uses Support Vector Machines (SVM's) to classify these audio segments into different groups. MFCC's and PLP's features where used to as input features, where the mean of each frame is computed and passed as the input feature to train the SVM model. A MIDI synthesized drum sound is then used as the replacement once the groups are classified and thus an effective replacement of a drum vocal imitation to MIDI synthesized drum samples are performed without any manual intervention.

1. INTRODUCTION

As the trend of modern smart devices is moving towards livelier user interfaces, it is equally important to build systems for musicians to have more interactive music production platforms. Taking this as a motivation this paper details about how beatboxing (Vocal Imitation of drums and percussive instrument sounds) can be converted into actual drum sounds. This system might provide a higher degree of freedom for the musician to compose more interesting musical rhythmic arrangements. There have been many recent works regarding the same using neural networks to convert beatboxing to drum sounds [1]. But those require more data and also more training time but using SVM, an efficient model could be built with equal or better accuracy than using a neural network. This model uses libsvm [2] a MATLAB executable C and C++ library for multiclass classification using SVM's. The section 2 gives the overall idea of the method used to convert beatbox to drums system, section 3 explains audio segmentation using spectral flux approach to detect onsets, section 4 talks about the how the dataset is produced and feature extraction procedure and in the final section we will discuss about the results and what are the future scope for the proposed system.

2. METHOD

The basic flow graph of the system is shown in the Fig.1. Firstly, recorded audio wave form is processed to estimate the onset, spectral flux method with a Gamma threshold of 0.2 and the onsets are extracted by employing peak picking technique with a threshold of 01. Using the onsets, the audio waveform is segmented into individual slices as each slice of audio represents an instance of vocal imitation of different drum sounds. And once the onset frames are known the frames are converted into samples and the audio segmentation task is performed.

The segmented audio is then passed to the feature extraction module, where the Mel frequency Cepstral Coefficients (MFCC's) and the perceptual linear prediction coefficients (PLP's) is computed and their mean value in each frame is calculated and all these coefficients are combined to form a single long feature vector for each segmented audio respectively.

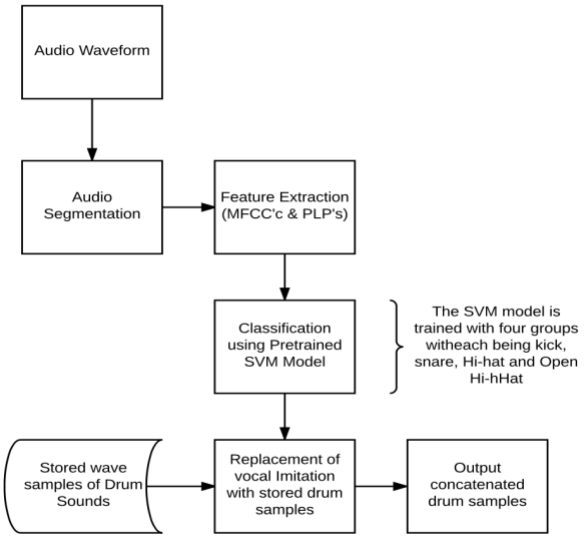

Fig.1. Flow graph to convert beatbox to drums

As observed form the flow graph Fig.1, the features are then passed to the SVM model. Where the SVM was pre-modelled with training set containing a short segment of vocal imitations of different drum sounds each with specific group label.

Using this SVM model the extracted audio segments are classified, that is determining which group they belong based on the similarities in the features in the training data and the segmented data. And the classified segments are replaced with the MIDI synthesized drum sounds that has been already stored in the system.

3. AUDIO SEGMENTATION

This section details the process of onset detection using spectral flux method and audio wave form segmentation using onsets and how these onset frames are then converted back to time and samples to separate the audio waveform.

3.1 Onset

Onset refers to the beginning of a musical note or other sound. It is mainly from the idea of a transient, all musical notes have an onset, but do not necessarily include an initial transient. In other words, transient can be thought as a short duration with high amplitude within which signal modulates at a faster rate. Fig.2, shows the onset and transient of a single instance of piano audio waveform.

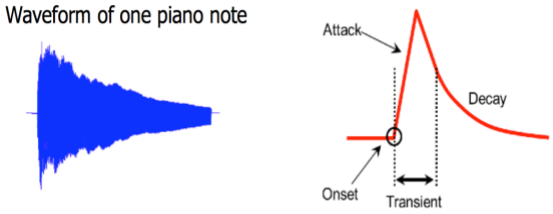

Fig.2 Waveform of one piano note and its envelope

Similarly, the onsets for the vocal imitation of audio segments are processed to detect the onsets of each instance.

3.1 Onset detection

Onset detection functions usually have low sampling rate compared to audio signals; thus, they achieve a high level of data reduction while preserving necessary information about the onsets. An onset detection involves distinguishing between the various types of change such as onsets, offsets, vibrato, amplitude modulation and noise. If an audio signal is observed in the time-frequency plane, the onset of a new sound has noticeable energy in the time-frequency bands in which the sound has noticeable energy (or amplitude) within some frequency bands is a simple indicator of an onset.

Alternatively, if we consider the phase of the signal in various frequency bands. It is unlikely that the frequency components of the new sound are in phase with various sounds, so irregularities in the phase of various frequency components can also indicate the presence of an onset. Further, the phase and energy (or magnitude) can be combined in various ways to produce more complex onset detection function. Onset detection could be achieved using spectral flux, phase deviation and complex domain methods. The section 3.2 explains the onset detection using spectral flux method.

3.2 Spectral Flux

Spectral Flux measures the change in magnitude in each frequency bin, and if this is restricted to the positive changes and summed across all frequency bins, it gives the onset function SF [3]:

$$SF(n) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|X(n,k)| - |X(n-1,k)|)$$

where H(x) is the half-wave rectifier function. Empirical tests favored the use of the L_1 norm here over the L_2 norm used [4][5] and the linear magnitude over the logarithmic (relative or normalized) function proposed by Klapuri [6]. Fig.3, shows the onset extraction from an audio waveform using spectral flux method.

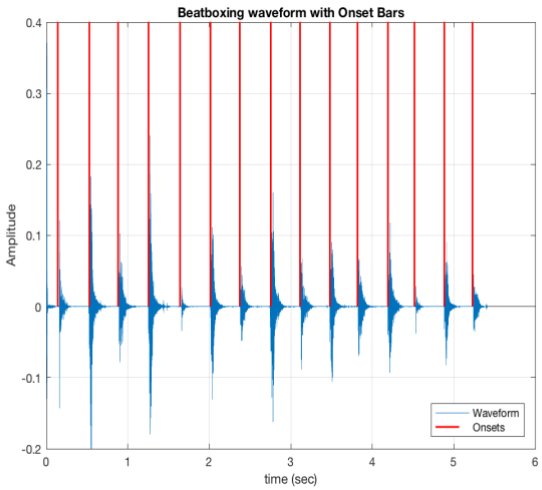

Fig.3 Onset detection using spectral flux method

As we could observe from the Fig.3 that the onset detection using spectral flux method mostly detects the exact onset frame and it can further be improved by adjusting the threshold parameters such as Gamma when calculating the and the by adjusting threshold for peak picking based on the style of beatboxing that is being processed.

3.3 segmentation

Once the onsets of the audio frame are computed we have the frame indices at which onsets occur. These frame numbers are then converted back to samples. That is by multiplying the frame number by hop size will give the onset in terms of samples in the audio waveform, one thing to notice is that the hop size should be the same as the hop size used to compute the spectral flux.

After converting the onset frames into their respective onset sample number. The sample points in between two successive frames are taken as one instance of vocal imitation of a drum sound. By doing so we get all the segments of audio based on the detected onset.

4. DATASET AND FEATURE EXTRACTION

4.1 Dataset

For training the SVM, I made my own data set using Audacity. The data set is of four groups Kick, Hi-Hat, Snare, and Open-HiHat. Each group has about 100 examples. The recordings where sampled at 22050 with bit depth of 16-bit and one second long. Since the dataset is small the application of the proposed model is constrained to work only when tested in my own voice and it might not work the best when tested with other's voice. Also, to make it robust to the segmented audio waveform. I extended the data sets with few training observations that are the segmented audio slices itself and the group labels to those are manually annotated.

4.2 Feature Extraction

As MFCC's are good in capturing the acoustic features of audio and speech efficiently. The segmented audio frame is processed to obtain the Mel-frequency coefficients. MFC is a representation of the short-term power spectrum of a sound. The MFC is converted into a single feature vector by taking the mean on each frame of the computed MFCC's to form a single feature vector of the training audio data as well as the test data which is the segmented audio. Fig.4 shows the MFCC's for three main classes kick drum, snare drum and hi-hat.

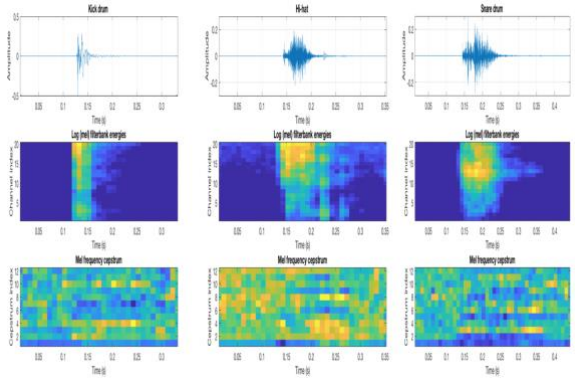

Fig. 4 MFCC's and filter-bank energies of each classes

In addition to MFCC's, PLP (Perceptual Linear Prediction) where are also extracted from the audio as PLP's are more robust to the differences in the acoustical features in between training data and the test data. So, the computed PLP's are also similarly converted into one single vector by taking the mean over each frame of the PLP matrices. And, Finally the MFCC's and PLP's feature vector coefficients are combined to form a single feature vector of size 45. The MFFC's and PLP's are computed based on the Ellis05-rastamat by Daniel P.W. Ellis [7].

5. SUPPORT VECTOR MACHINES

SVM's are supervised learning models associated learning algorithms that analyzes data used for classification and regression analysis [8]. SVM could be used as a binary classification model where there is only two set of groups to classify and also it could be used as a multiclass classification problems where the algorithm kind of groups into cluster based on one vs all approach. For a set of training examples each with labels denoting to which group they belong to, SVM training algorithm builds a model that assigns new examples to one category or other.

In the proposed system, as mentioned before there are four groups and they are labeled as shown in Table 1:

Instrument Name	**labels**
Hi-Hat	1
Kick Drum	2
Open HI-Hat	3
Snare Drum	4

Table 1. Group name and labels

By concatenating the corresponding labels with the dataset, the dataset is preprocessed and ready to be used for training and testing the SVM model, 80% of the dataset was used to train the SVM model to fit the data. And, the remaining 20% was used to test the obtained model from the SVM.

Fig.5 shows minimum objective and estimated minimum objective of the trained SVM model.

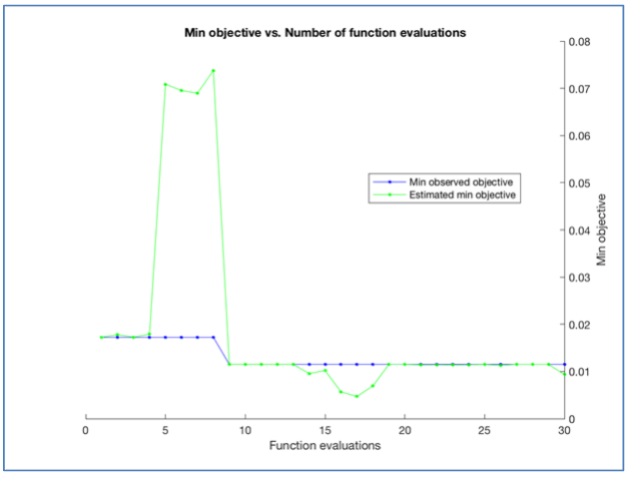

Fig.5 Minimum objective Vs estimated objective

6.RESULTS

The trained SVM model achieved an accuracy of about 95.7% in the test data. And achieved a class accuracy of about 70% and when tested with segmented audio the accuracy was about 85% Which is almost in par with results obtained through neural networks model designed by Devansh Zurale et.al [1]. Table 2 shows the accuracy comparisons of two model.

Model	Accuracy
DNN	96%

SVM	95.7%

Table 2. DNN vs SVM Accuracy

As we could observe that the results are promising with both the methods, but one advantage of using an SVM instead of Deep Neural Networks is SVM takes much lesser time to train the model where as DNN's take longer time when compared to SVM to train and optimize the weights to classify the labels.

7.DRUMS SYNTHESIS

Once the SVM model classifies each slices of the segmented audio, based on the classification the slices are replaced with their respective MIDI synthesized drum sounds. The length of each synthesized audio is same as the size of the segmented audio. The choice of MIDI samples is totally up to the preference of the user. And finally, all the synthesized samples are concatenated to form one single audio waveform of Drums.

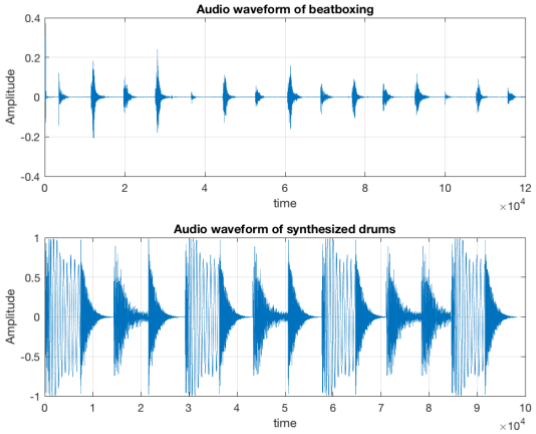

Fig.6 Audio waveform of beatboxing and synthesized Drums

8.DISCUSSIONS

Several methods were tested for onset detection but the spectral flux method produced higher accuracy and was easier to implement. And, at times the model does not perform that well on the segmented audio and I feel the reason for that is because of the smaller data size. By increasing the data size the system should be more effective than its performance at the present.

And in order to decrease the dimension of the input features to the SVM model, PCA feature reduction was done on the features but it did not have any positive impact on the system. Hence the final model did not have PCA in it.

9.CONCLUSIONAND FUTURE WORK

In this paper, we discussed a SVM model which achieved a successful conversion of the input beatboxing audio to an MIDI synthesized drum recording with an accuracy of about 95% is developed. The system accuracy however, is severely affected in the presence of noise. Which has to be addressed in future in order to make this system to be more robust to all environments.

There could be various future extensions to the proposed system such as adding different classes to classify (i.e., having more different drum sounds as vocal imitation). And by extending the dataset with various other human observations this system could be a potentially used as a user independent model.

10.REFERENCES

[1] Devansh Zurale and Jonathan Michelson: "Beatbox-to-Drum Conversion," *Carnegie Mellon University.*

[2] Chang, Chin-Chung and Lin, Chih-Jen: "LIBSVM," *ACM Transactions on Intelligent Systems and Technology,* Vol. 2, issue.3, pp. 27:1__27:27, 2010.

[3] P. Masri, "Computer modelling of sound for transformation and synthesis of musical signal," Ph.D. dissertation, Univer- sity of Bristol, UK, 1996.

[4] C.Duxbury,M.Sandler,andM.Davies,"Ahybridappro ach to musical note onset detection," in Proc. Int. Conf. on Dig- ital Audio Effects (DAFx-02), Hamburg, Germany, 2002, pp. 33–38.

[5] S.Dixon,"Learningtodetectonsetsofacousticpianoton es," in Proc. MOSART Workshop on Current Dir. in Computer Music Res., IUA-UPF, Barcelona, 2001, pp. 147–151.

[6] A. Klapuri, "Sound onset detection by applying psychoa- coustic knowledge," in Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc., Phoenix, Arizona, 1999.

[7] Chang, Chih-Chung and Lin, Chin-Jen, LIBSVM: A Library for Support Vector Machines, ACM Transactions on Intelligent Systems and Technology, volume 2 issue 3, 2011.

[8] Mingxia Liu, Daoqiang Zhang, Songcan Chen, Hui Xue, "Joint Binary Classifier Learning for ECOC-Based Multi-Class Classification", Pattern Analysis and Machine Intelligence IEEE Transactions on, vol. 38, pp. 2335-2341, 2016, ISSN 0162-8828.