Journal of Computer Science

Original Research Paper

Beatbox Classification to Distinguish User Experiences Using Machine Learning Approaches

[1]Jason Martanto and [2,3]Iman Herwidiana Kartowisastro

[1]Computer Science Department, Bina Nusantara University, Jakarta, Indonesia
[2]Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[3]Computer Engineering Department, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia

Article history
Received: 28-07-2024
Revised: 04-12-2024
Accepted: 24-12-2024

Corresponding Author:
Iman Herwidiana Kartowisastro
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: ihkartowisastro@binus.ac.id

Abstract: Research regarding beatbox classification has generated a relatively significant growth in the past decade. Although the differences between contributors' expertise within a vocal percussion dataset have been mentioned in previous works, the impact of those discrepancies has not been thoroughly investigated. In this study, the authors explore performances of machine learning algorithms for beatbox classification, with an emphasis on prior beatboxing experience affecting dataset. Throughout this study, feature extraction is conducted by the use of 4 methods, i.e. Spectral Centroid, Spectral Magnitude, Spectral Contrast, and MFCC, while machine learning method to perform classification is through the use of KNN (3,5,7), Adaboost, LSVM one-vs-one, LSVM one-vs-rest, SVM one-vs-one, SVM one-vs-rest. This study shows that performing a beatbox classification requires more thought into the differences between the skill level of the dataset (inexperienced and trained/professional). Points of concern include the shorter time span in a trained beatbox dataset to segment and classify before the next onset begins, in which some sounds were even found to be smaller than 0.01 ms. For classification experiments using several feature extraction techniques and machine learning models, experiment results show that MFCC (n_mfcc = 22) delivers the best feature representation for our KNN, multi-class and non-linear SVM classification model.

Keywords: Vocal Percussion, Machine Learning, Classification, Beatbox Experience

Introduction

Artificial intelligence has shown its powerful capabilities in our daily life. Many artificial intelligence based products could be seen in so many fields, including musical note transcription, from drum (percussive instruments) transcription (Wu *et al*., 2018) to a more varied music audio transcription (Saputra *et al*., 2021). Youths in 1970's found another way of creating percussion sound through the use of human vocal.

Vocal percussion itself, known by most as beatboxing, is the art of creating a rhythmic and melodic imitation of percussive musical instruments such as a drum kit. Beatboxing originally started as a unique accompaniment to hip-hop in the 70's (humanbeatbox.com, 2005). Nowadays, instead of the traditional drumkit imitation, beatboxers have improved their techniques to be as technical and unique than ever before. Not only do they have a more diverse and bigger repertoire of unique sounds, they also start reflecting more modern forms of music into their performances (Blaylock, 2023). This turned beatboxing from a quirky backing track into a standalone form of music that explores the limits of human skill in producing voice.

Beatbox classification is a sub-field in Music Information Retrieval (MIR) that deals with the grouping of vocal percussion sounds to their proper respective labels. One of the biggest potentials in beatbox percussion is the fast and spontaneous method of input for rhythm in music. In order to assist with the public learning how to beatbox, a program which implements effective beatbox audio recognition might provide an easy and fun environment to develop their sounds properly. For example, implementing beatbox

classification in a beatboxing video game, to disguise the act of learning into a fun activity (Kylie *et al*., 2011), such as the case with the singing video game One Hand Clapping (One Hand Clapping, 2020).

Other than their usage in artistic performances, beatboxing can also be utilized in other fields such as its usage as a form of speech therapy for the intellectually impaired (Icht, 2019) and a tool for musical rhythm education for the youth (Matveeva *et al*., 2019). Although field of beatbox classification research is not exactly unexplored, it has not seen as much attention as other audio related sub-fields such as speech recognition (Malik *et al*., 2021) and even MIR fields like drum transcription (Wu *et al*., 2018). This does impact the field in its lack of publicly available datasets within the field.

An emerging from the beatbox classification space is that classification tasks that are carried out on dataset that only represents a subset of the public based on their skill level. Vocal percussion datasets that are employed in these research are usually either focused on experienced (Stowell and Plumbley, 2010) or a complete amateur to the artform (Delgado *et al*., 2019). Although there have been previous research (Sinyor *et al*., 2005) that has touched upon this topic, they didn't conduct a deeper dive into the matter. This study is created to explore the subject of beatboxing skill level and infer their performances for a user-agnostic approach on beatbox classification regardless of prior experience. This research will be initiated by discussing the fundamental technical aspects of beatboxing and the discrepancies between the different skill levels. The experiments are then conducted on a variety machine learning models using several well-known feature extraction methods within the surrounding field of research. Lastly, the performance of the models will be collected and evaluated using the f1 score extracted from multiple different models.

Beatboxing Technique

For the sake of clarity in this study, the level of skill level of a dataset will always be explicitly stated to avoid confusion. As this research only focuses on the percussive aspect of beatboxing, the term "beatbox" and "vocal percussion" are used interchangeably as they represent same concept within this research.

Although techniques in beatboxing have advanced significantly, replication of percussive instruments remains the most fundamental form of the art form. It's widely accepted in the beatbox community that beatboxing has 3 basic primary sounds: kick {b}, hi-hat {t}, and snare{k}. Notation for beatbox sounds can be done in curly braces as stated in (Revd Gavin and Mark, 2014).

Additionally, we can also annotate the beatbox sounds phonetically according to the International Phonetic Alphabet (IPA) which can later help with Language Model (LM) approaches for beatbox classification. The 3 basic sounds are usually the farthest extent of the general public's knowledge regarding beatboxing techniques. However, even those basic techniques are perceived differently by trained/ professional beatboxers in terms of their actual depth and complexity.

What the public refers to as a snare sound is usually the technique known as the k-snare and it's phonetically made using a voiceless velar plosive or [k] in its IPA format. Even the k-snare itself has multiple variations depending on the breathing direction and tongue position such as the inward k-snare (made by breathing in). The k-snare is only one of the many varieties of snare sounds such as the pf and spit snare. This is why the snare sound is often considered to be most versatile sound to learn among the three basic techniques.

Another example of depth in basic beatboxing techniques can be found in the hi-hat sound. The hi-hat sound can be phonetically described to be made by a voiceless alveolar sibilant affricate [t͡s]. Trained/ professional beatboxers can extend the tail of the uttered sound with a hiss to replicate an open hi-hat or stop it shortly after the initial attack to replicate a closed hi-hat. On top of that, the characteristics of the hi-hat might also differ according to how the sound is vocalized. For example, uttering a plosive unvocalized "t" [t͡s] and "ch" [tʃ] are both valid beatboxing techniques for replicating a closed hi-hat sound.

In beatboxing, the variations of kick sounds are rather limited compared to the other basic sounds. The standard technique for creating a kick sound is uttering a sharp and loud voiceless biblical plosive [p]. Although other kinds of kick sound do exist in beatboxing (lip roll kick, throat kick, etc.), they are more scarcely used compared to the standard kick technique.

Proficiency in Beatboxing

An amateur vocal percussion dataset may be applicable for recognizing amateur attempts at imitating a percussion sound. However, they might not be as viable to train models on recognizing intermediate to the professional level of beatboxing. This incompatibility will be even more prominent when the roles are reversed. These differences of characteristics can be seen in the following components of beatboxing:

- Technique: Amateur vocal percussion is usually made as a spontaneous attempt at replicating a percussive sound without any prior knowledge of the proper method, or a short and somewhat plosive utterance of letters that has a similar characteristic to the beatbox sound
- Consistency: Regardless of how the technique is performed, one of the most distinctive characteristics

of a trained/ professional beatboxer is consistency. To have a mastery of a beatboxing technique, a beatboxer must be able to consistently replicate the loudness, pitch, and the general timbre of the sound. After all, a kick drum in an actual drum set will create the same sound consistently. This level of consistency is usually not found in amateur attempts at vocal percussion unless they're using a vocalized phrase to replicate beatbox sounds

Audio Machine Learning

One of the most distinct differences when applying machine learning to audio-related tasks can be found in the dataset, how it's represented, and their features. In most cases, the initial step involves selecting the most suitable representation of audio signal for our task. This requires knowledge in the field of Digital Signal Processing (DSP) which is concerned with the manipulation and analysis of signals that have been digitized. After selecting the most suitable representation, we can proceed with extracting task-specific features to train our model. Most research in audio machine learning space usually deals with speech and music as their main topic of discussion.

Automatic Speech Recognition (ASR) is the task of understanding and interpreting real spoken human speech. On the other hand, Music Information Retrieval (MIR) is the research of extracting important information from music signals. Research fields such as Automatic Drum Transcription (ADT) and beatbox classification fall into the field of MIR. However, an important point of thinking is that vocal percussion is performed with the human mouth, which is a major point of concern in speech recognition research. On the other hand, beatboxing is also concerned with the percussive component of the sounds which is the point of focus in the field of ADT. Thus, both research fields have a logical connection with things concerning beatbox classification.

Automatic Drum Transcription (ADT)

Beatbox is the replication of percussive instruments such as a drumkit. We can thereby draw a connection between the characteristics of a drum kit and beatboxing. As papers involving beatbox classification are less abundant than drum classification, it might be possible to reinforce our understanding on beatbox classification further by employing previous ADT research.

An important overview of ADT research was meticulously written by a group of researchers within the field (Wu *et al*., 2018). This comprehensive analysis was able to present the growing trend within the ADT field of research until the year 2017. The paper showed that there was a large chunk of research within the field that utilized Short Time Fourier Transform (STFT) (Cahyaningtyas *et al*., 2023) and its Low-Level Features (LLF) to represent the spectral features of percussive sounds. Additionally, machine learning methods such as Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) seems to be one of the most used machine learning methods used to classify sound events within the field.

Short Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) is a technique used to process continuous signal into a time-frequency domain for further analysis purposes. Essentially, it's done by observing the frequency composition of a signal in a short time segment. This addresses the problem that a traditional Fourier transform unable to accomplish, like the incapability of time-varying analysis. Nowadays, a lot the of modern spectral feature extraction techniques are derived from STFT like spectral magnitude, spectral centroid, and others.

The process is initiated by chopping a signal into smaller segments. A windowing function is applied to the segmentations to help with the problem of spectral leakage, which can cause the increase of inaccuracy in the performance of the models that utilizes the features extracted. The result of STFT is a time-frequency spectrogram, where time is represented along the x-axis, frequency along the y-axis, and the intensity reflects the amplitude or power of the signal at various frequencies and point of time

Automatic Speech Recognition (ASR)

The field of speech recognition is concerned with a machine's ability to be able to recognize, transcribe and understand human speech. Mel Frequency Cepstrum Coefficients (MFCC) (Ranjan and Thakur, 2019) has been used in the field of ASR and has shown a high level of success in multiple different speech recognition cases. It has even shown success as a feature extraction method in languages outside of English (Rynjah *et al*., 2022). According to a literature review on speech recognition (Malik *et al*., 2021), MFCC was concluded to be one of the most commonly used feature extraction technique in the case of ASR. The review also stated that the use of Hidden Markov Model (HMM) in the field is very common but was concluded to not be as optimal as other methods like SVM.

Mel Frequency Cepstrum Coefficient (MFCC)

In the audio field, especially speech recognition, MFCC has been utilized to successfully extract features from speech, and music (Haq *et al*., 2020; Ranjan and Thakur, 2019; Tiwari, 2010). MFCC is a feature that is extracted from the 'Mel-scale', which is a scale that is used to represent the subjective pitch of a human ear. Essentially, a linear difference in frequency does not subjectively sound like a linear pitch change to human

hearing. Because of this, the Mel-scale was invented as a linear scale to represent the subjectiveness of a note change, even though they are not linear when observed frequency-wise. Since we as humans can listen to a note change subjectively, being able to have computers perceive the change in pitch like human hearing is valuable. To produce MFCC, we need to convert the frequency-domain spectrum of our signal into the mel-scale. A triangular filter-bank that is designed to replicate human hearing is applied to create the mel-frequency spectrum. Converting a normal frequency into a mel-scale frequency require the following mathematical formula:

$$m = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Here in eq. 1, m is the result of the mel-scale conversion and f is the frequency of the input audio signal. To convert the generated mel-frequency spectrum into the form of 'cepstrum', the use of Discrete Cosine Transform (DCT) is employed to convert the spectrum back into the time-domain. Additionally, DCT returns real-numbers coefficients instead of complex numbers, which is easier to utilize for practical purposes. In real use cases, only the first few numbers of coefficients will be used to represent features of an audio signal. The number of coefficients to choose depends on the use cases of the MFCC. For example, ASR have found success in utilizing around 12-13 coefficients to represent speech features (Haq *et al*., 2020).

Related Works

One of the first research in beatbox sound classification was done by Sinyor *et al.* (2005). Participants contributing to this research included 3 beatboxers and 3 non-beatboxers. In total, the dataset consists of 1200 samples, which they segmented and annotated manually on the software Audacity. Although the discrepancies between beatbox skill level was mentioned and showed in their waveform, no further exploration was done into the topic. Utilizing the dataset created, they were able to classify beatbox sounds into 5 different classes with an accuracy of 95.5% using the Adaboost ensemble learning method with C4.5 decision tree as the base learners. Cutting down the number of classes in the model to 3 instead of the original 5 increased the accuracy even further to 98.15%. In this case, reducing the number of classes might not necessarily be a loss as they were still able to classify the 3 most fundamental sounds of beatboxing (bass, hi-hat, and snare).

A research by Picart *et al.* (2015) explored utilized Hidden Markov Model (HMM) as the classification method for beatboxing. They recorded a beatbox dataset that consisted of 2 trained/ professional beatboxers and did their classifications on both the percussive and instrumental aspects of beatboxing. A total of 5 percussive classes and 9 instrumental classes were identified to be the target of classification. Using MFCC as their features extraction method, they were able to reach the best results on the percussion recognition using 22 coefficients with an error rate of 9%. On the other hand, the instrumental recognition was only able to reach an error rate of 41% at its highest after employing 18 coefficients.

To do real-time beatbox performance recognition, a computer must do the recognition process and trigger the wanted action from the output of the recognition. However, this brought upon the question of latency and the ideal delay duration between recognizing the sound and outputting the result. This was answered by another research (Stowell and Plumbley, 2010) when they examined how the time between the detection of an onset and the classification task affects the performance of a naïve bayes classifier model. For their research, they collected a dataset which consisted of 14 recordings by different beatboxers and named it "beatboxset1". They concluded that a delay time of 23 milliseconds was the ideal value for their dataset classification.

To assist in the recognition process for amateur beatboxing, a research paper (Delgado *et al.*, 2019) has provided an amateur vocal percussion dataset that is open to the public. This dataset consists of 9780 annotated utterances of vocal percussion made by 28 untrained participants. For each contributor, there are 5 pairs of audio recordings and annotation files. Four of these file pairs correspond to the recording and labeling of kicks, snares, closed hi-hats, and open hi-hats, while the last remaining pair is assigned to an improvisation by the participant. The significance of this dataset lies in its representation of the public majority that has no prior experience in beatboxing (no expertise). They also carried out an onset detection experiment using the dataset which concluded that DSP features were able to outperform deep learning methods.

A toolkit that is usually used for automatic speech recognition named Kaldi, was adapted to create a beatbox sound recognition system by Evain *et al.* (2021). This experiment was rather successful at implementing Kaldi for beatbox sound recognition. They also coined the term 'boxemes' which is short for beatbox and phonemes to define each individual phoneme representation of beatboxing sounds. The experiment was conducted with a dataset that was recorded by 2 trained/ professional beatboxers with differing skills. Within their work, the utility of 13 or 22 MFCC, 13 PLP, and 40 Bank was tested as their selected feature representations. They were able to lower the boxeme error rate to approximately 13,6% on the GMM-HMM model using MFCC with 22 coefficients.

Another research (Delgado *et al*., 2022) has explored a different way of representing beatbox sounds by creating supervised CNN embedding models to create feature sets with different levels of abstractions. They tested syllable-level, instrument-level, and phoneme-level beatbox sound annotations and compared the results with the baseline methods that they defined. After testing the accuracies of the embedding with a KNN classifier, they discovered that using a syllable-level annotation seems to have resulted in the best performance with an accuracy of approximately 87.4%.

Ramires (Ramires, 2017) researched the potential of an automatic vocal percussion transcription that turns beatbox signals into the form of MIDI inside of the Digital Audio Workstation (DAW) named Ableton. He employed the Sequential Forwards Selection (SFS) algorithm as a feature selection method and KNN as the machine learning model of choice. The dataset used in this research was recorded by Ramires with 11 men and 9 women as the participants, in which only a single contributor has a beatboxing skill. The impact of different recording qualities was also observed and showed that the recordings with a laptop microphone underperformed significantly compared to the studio microphone (AKG c4000b) and an Ipad microphone. As an example, the F-measure result for the kick sound using the LVT system was able to achieve a high 91.4% on the AKG microphone, while the laptop microphone was only able to reach a terrible score of 27.9%.

When dealing with small datasets, applying audio data augmentation might be the most optimal solution when new samples are hard to come by. A paper by Wei *et al*. (2020) compared a number of audio augmentation techniques such as pitch-shifting, noise injection, and others showed an improvement in performance. They also proposed a new method called Mixed Frequency Masking which showed the most improvement compared to other tested augmentation techniques.

Although research in beatbox machine learning has been on the rise in the last 2 decades, the authors noticed the common trend of employing datasets that only represent either amateur or professional beatboxing as seen in Table (1). This absence of discussion appears to be a gap within the field that needs to be addressed, thus making it this research's purpose to explore and document.

Table 1: Expertise level of existing vocal percussion works

Research	Year	Expertise
Sinyor *et al*. (2005)	2005	Trained and Amateur
Stowell and Plumbley (2010)	2010	Trained
Picart *et al*. (2015)	2015	Trained
Ramires (2017)	2017	Amateur
Delgado *et al*. (2019)	2019	Amateur
Evain *et al*. (2021)	2021	Trained and Amateur

Materials and Methods

Methodology of experiment in this research will be arrange in the following order: Dataset augmentation, analysis, processing; testing feature extraction methods; and finally evaluating the performances of the classifier models. Scrutinous observations of the dataset innate characteristics and their impact in the context of machine learning will be carried out in the experimentation section.

Publicly available datasets within the beatbox research field are selected for training and testing our models. The Amateur Vocal Percussion (AVP) (Delgado *et al*., 2019) dataset was chosen to represent individuals with little to no experience of beatboxing while beatboxset1 (Stowell and Plumbley, 2010) (referred to "BTX") was chosen to represent the demographic with prior training.

Since there are discrepancies between the labelling format of each dataset, annotations of dataset will be reduced and grouped into a percussive sound that they represent together. Since the AVP dataset has the least number of labels, its annotation format was selected for our experimentation. The 4 classes chosen as the target of classification are open hi-hats (hho), closed hi-hats (hhc), kick, and snare.

Segmentation of beatbox sounds will be carried out according to the annotation files that came with each dataset. This will be implemented using Pydub 0.25.1 and the cutting will be done from the onset time specified by the annotation file to the next. The result of which can be seen on Table (2). Furthermore, to tackle the problem of a relatively small and imbalanced number of the input data (especially the hho of BTX), an audio dataset augmentation technique was applied to increase training data and improve the variability of the dataset. This was done because audio augmentation has been proven to improve the performance of classification tasks (Wei *et al*., 2020).

Implementing the augmentations was be done by using the Audiomentations (Jordal *et al*., 2023) library. To augment the data, both pitch shifting (semitone from -1.5 to 1.5), gaussian noise injection (amplitude from 0.001 to 0.003), time stretching (slowed by 0-15%) were applied to the AVP dataset. The BTX dataset was only augmented using pitch shifting similar to the one applied to AVP because the dataset was observed to already contain a lot of noise within the recordings. A mixed dataset was constructed from the features of the 2 datasets and was used to see how well a dataset from both amateur and professional contributors can be utilized to train a classifier model for both types of datasets.

Table 2: Result of segmentations for both the AVP and BTX datasets according to the manual onset-detection

Dataset	Kick	Snare	HHO	HHC
AVP	1064	1002	967	1047
BTX	996	972	271	957

Amateur Dataset Processing

A total of 28 participants was recorded for the AVP dataset and each of them contributed 5 pairs of audio and annotation file. As none of the participants in this dataset was trained in beatboxing, they attempted to replicate beatboxing techniques by uttering short single syllable phrases (Fig. 1). For instance, the kick sound is usually attempted by uttering the phrase "pem" [pəm]. The annotation format of this dataset consists of the 4 different sound classes that we have defined to be the target for classification. Certain data points that were wrongly identified as onsets by the annotation file has been removed. Additionally, segments that were shorter than 0.011 ms (512 sample / 44100 sample rate) were skimmed off from the dataset.

Experienced Dataset Processing

The BTX dataset was selected to represent the experienced beatboxer demographic within the public. It included battle clips and voluntary submission by 14 different trained/ professional beatboxers. Manual onset detection and labelling in this dataset was done by Helena du Toit and Diako Rasoul. For our experiments, we arbitrarily chose the one done by Helena. The labelling of this dataset was filled with many variations and was reduced to the same 4 classes in AVP as such: kick ("kd"), snare ("s", "sb", "sk"), hho ("ho"), and hhc ("hc").

One of the significant problems to be observed in the dataset is the resolution of samples to be analysed. Some data point in this dataset struggled with being too short to be inputted to the current STFT resolution and this was caused by the speed that a professional beatboxer can achieve when alternating sound amplitudes to mimic percussive sound (Fig. 2). This problem is remedied by padding silence samples (until 2048 long) at the end of an audio segment.

Feature Extraction

To extract the features inputted into our classification experiments, we have employed MFCC, and STFT derived techniques, i.e. spectral magnitude, spectral centroid, spectral contrast. These features were extracted using the Librosa library, which is an easy-to-implement and resourceful tool for audio signal related research. As they were all derived from STFT, the parameters used in the experiment will be similar across the different feature extraction techniques. These parameters include a sample rate of 44100 Hz; the use of 'Hann' window; window size and n_fft of 512; hop length of 256. Unique to MFCC, an additional parameter to control the number of generated coefficients was set to 22, as it has shown success in previous works (Evain *et al*., 2021; Picart *et al*., 2015).

A simple randomized train test split is then applied to the extracted features to randomize and split the data into 2 subsets for training and testing purposes using a ratio of 8:2.

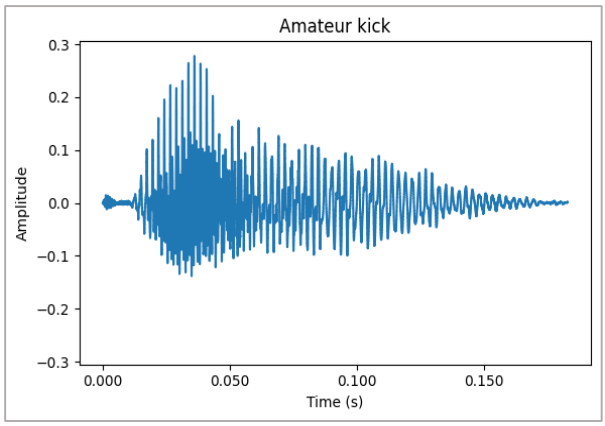

Fig. 1: Waveform of an amateur vocal percussion kick

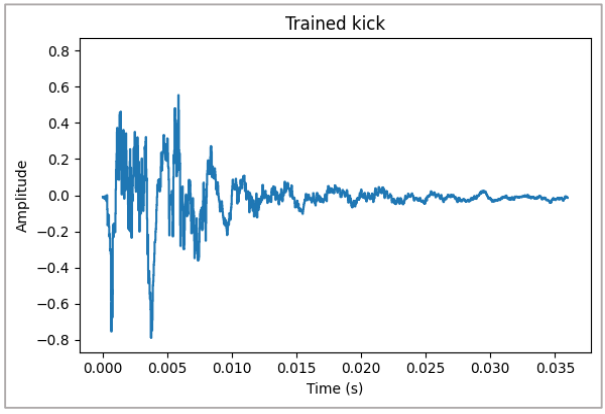

Fig. 2: Waveform of a professional vocal percussion kick

Classification Model

The machine learning algorithms selected for our classification experiments include KNN and both linear and non-linear SVMs, due to their success with ADT research (Wu *et al*., 2018). Three different values were also tested on the parameter n_neighbors to ensure we tested the highest possible performance. For the SVM models, we set the penalization to 'l2' and loss function to 'square-hinge'. Due to the nature of the data, since SVM cannot accomplish a multi-class classification, we applied the one-vs-one and one-vs-rest method to our both of our linear and non-linear SVM models.

Lastly, we also tested the ensemble method Adaboost which utilized 300 decision trees (max_depth = 1) at its base estimator and a learning rate of 0.5 similar to the best performing model of a previous research (Sinyor *et al*., 2005). The performances of the models were evaluated using the f1 score. These machine learning models and evaluation method were implemented using the python library sklearn 1.2.2 for the ease of implementation.

Results

To get an overview on the results, a visualization using several grouped bar charts was made and can be seen in Fig. (3). The performances of each model were evaluated using their f1 score and were displayed separately for each individual class. Each of the coloured bars represent different classes: Orange represents open hi-hat (hho), green represents closed hi-hat (hhc), blue represents kick, and red represents snare.

As seen in the figures, both spectral contrast and spectral centroid were seen to perform very poorly in their overall performance across all models. Even its best performing class, which is the kick, was shown to be a horrible 0.60 F1 score. Spectral centroid was especially terrible when it comes to distinguishing between closed and open hi-hats, which makes it a very unsuitable feature extraction technique for a 4-way beatbox classification like our case. As it is usually used to measure the brightness of a sound, spectral centroid was observed to have better results in identifying kick sounds which are more prominent in its lower frequencies.

Spectral contrast showed a trend of a mediocre to low performance on 2 of the 4 classes (kick and snare) while underperforming for when it comes to representing the hi-hats (especially true for the closed hi-hat class). Once more, the kick was the label to have the best classification performance. Although it was still a low 0.71, which was achieved on the SVM one-vs-one model, it was still a massive step-up from the results that spectral centroid produced.

On the features side, MFCC with 22 number of coefficients has shown the best performances across the different feature extraction techniques that were tested. However, the features represented by spectral magnitude were still able to perform very well and is only slightly behind MFCC. Although MFCC performed with the highest performance on the majority of models, spectral magnitude still produced a marginally higher result on Adaboost, and both linear SVMs.

All of the KNN configurations (n_neighbors = 3, 5, 7) has consistently shown the highest result across all of the different feature representations. On the other hand, Adaboost and both multi-class linear SVMs seem to perform the worst across all of the different experiments. Both one-vs-one and one-vs-rest SVM produced results with a relatively miniscule difference between them, with the most prominent difference being their performance on classifying closed hi-hats. The one-vs-rest method for multi-class SVM (both linear and non-linear) performed ever so slightly lower compared to the one-vs-one method.

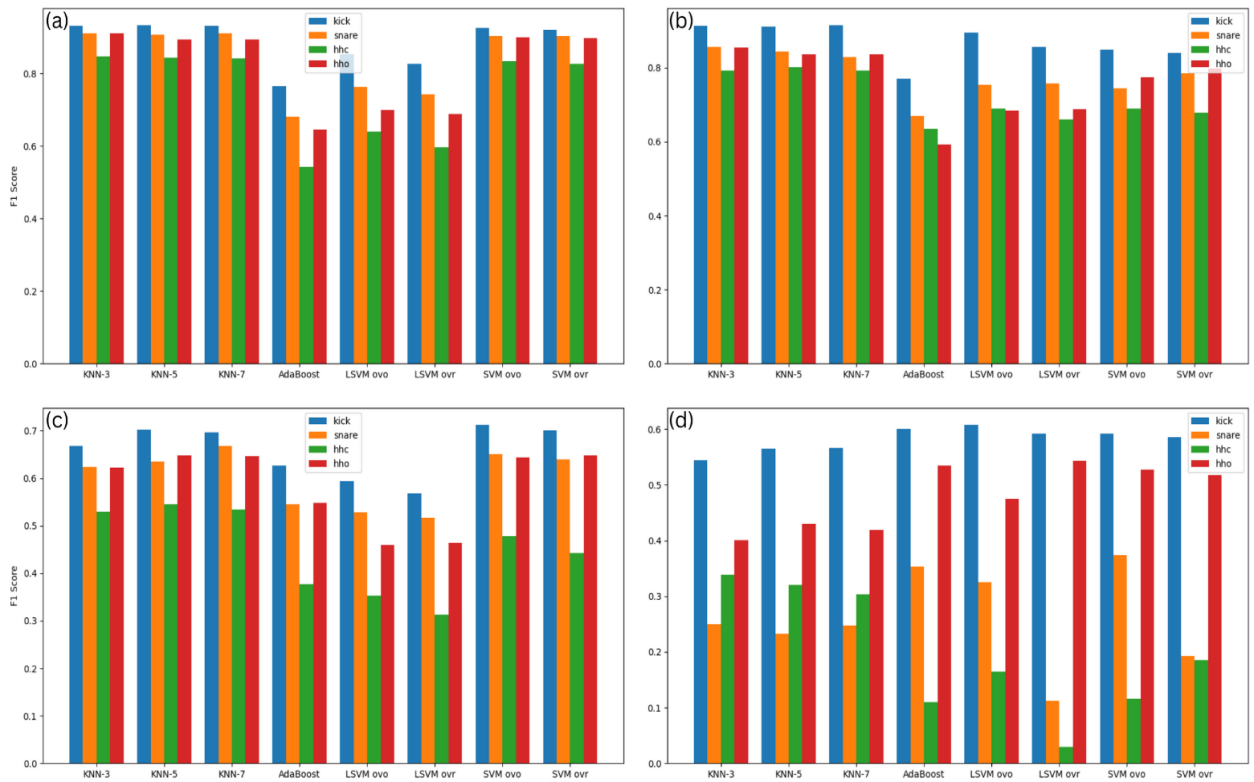

Fig. 3: The f1-scores of all 4 classes on several different machine learning models using several feature representations. (a) MFCC, (b) Spectral magnitude, (c) Spectral contrast, and (d) Spectral centroid

Discussion

An interesting observation of the features tested indicates that a mixed dataset appears to be better identified when using a feature more commonly used in identifying the human voice (MFCC), perhaps due to the nature of the amateur dataset within, compared to features more commonly used in identifying percussion sounds (Spectral Magnitude, Contrast, Centroid).This level of success with MFCC has been seen in previous work (Evain *et al*., 2021; Picart *et al*., 2015) and is now supported even further by the comparison done with other feature extraction techniques, the variety of machine learning methods tested, and the utilization of a more encapsulating dataset within this research.

Our worst performing combination of being LSVM o-v-r model and spectral centroid features as seen in the confusion matrix on Fig. (4). Seeing a lot of closed hi-hats predicted as open hi-hats was not that odd considering their similarities, but the fact that a lot of closed-hats was predicted as kicks was a little unexpected as they are usually very distinguishable with other features extraction methods. It appears that sounds with a "sharper" sound with mid to high frequency are mostly predicted to be an open hi-hat sound while the lower frequency sounds are mostly predicted as a kick, though there are still outliers to this observation.

The difference between KNN classes doesn't contribute to a significant improvement between them. Thus, suggesting that even the smallest neighbor configuration of 3 might be sufficient to identify beatbox sounds with a good performance, when given features derived from spectral magnitude and MFCC. Though whether it improves or degrades with more classes depends on the features.

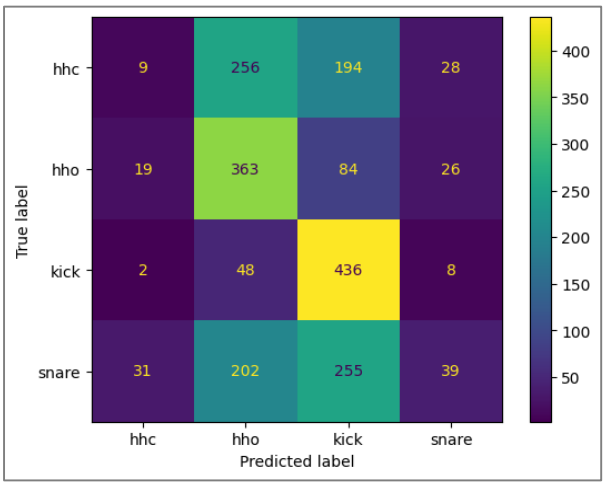

Fig. 4: Confusion matrix of worst performing model (Spectral centroid on LSVM o-v-r)

The drop-off in performance when using Adaboost with decision trees seems to suggest that the MFCC and other spectral features that were tested are not as compatible as to classify mixed beatbox sounds shown in previous work (Sinyor *et al*., 2005). Linear SVM seems to always fail to achieve a satisfactory result which suggest that linear boundaries are not as capable as the more complex boundaries brought upon by non-linear SVM when classifying multiple beatbox sounds regardless of features.

Conclusion

Beatbox performances can be carried out by anyone with differing level of skill. However, expertise level in beatboxing may affect how quick alternating sound amplitudes can be generated to mimic percussion sound. This manuscript highlighted a thorough exploration in beatbox classification and their performances when using a mixed dataset from both ends of the expertise spectrum. After testing several different models and feature extraction methods, MFCC with 22 coefficients was able to achieve the highest performance on a majority of the model tested. However, spectral magnitude was able to achieve a relatively more stable performance across models. It was also observed that KNN was able to perform consistently between all the models tested, although it can't be said to perform as well with features such as spectral centroid and contrast.

We are also acknowledging limitations within our work such as the manual on-set detection that is not ideal in a practical setting and the limited number of datasets that we have, which are things to investigate with future research. Many interesting issues in both music and beatbox recognition such as building classification techniques to distinguish music harmony (two or more notes heard simultaneously) and rhythm (the pattern of sound placements within a certain time) remain open for further work with the use STFT and MFCC approaches in extracting features. There is also the future task of achieving a live transcription of beatboxing sounds regardless of their skill level, which would be the main goal of this research field. With an accurate live transcription of beatbox sounds, practical applications such as a beatbox transcription directly to a MIDI file in a music producing software, and games to teach rhythm with beatboxing will be possible.

Acknowledgment

We acknowledge Dr. I Gede Putra Kusuma, the Head of Computer Science Department, Binus Graduate Program, Bina Nusantara University, for his motivation and support during the creation of this manuscript.

Funding Information

The authors would like to thank Bina Nusantara University for providing a professorship research program and funding opportunity. This study would probably not have been possible without such an opportunity.

Author's Contributions

Jason Martanto: Collected datasets, carried out literature study and experiments, created draft of the manuscript, and checked the manuscript in English.

Iman Herwidiana Kartowisastro: Supervised and assisted in planning out and conducting research analysis, and revised draft of the manuscript.

Ethics

This study was an original work and has been agreed upon by both authors to not involve any ethical issues.

References

Blaylock R. (2023). Why beatboxing? ICU Working Papers in Linguistics(ICUWPL), 24, 13–26. https://doi.org/10.34577/0002000007

Cahyaningtyas, Z. A., Purwitasari, D., & Fatichah, C. (2023). Deep Learning Approaches for Automatic Drum Transcription. EMITTER International Journal of Engineering Technology, 11(1), 21–34. https://doi.org/10.24003/emitter.v11i1.764

Delgado, A., Demirel, E., Subramanian, V., Saitis, C., & Sandler, M. (2022, April 10). Deep Embeddings for Robust User-Based Amateur Vocal Percussion Classification. https://doi.org/10.48550/arXiv.2204.04646

Delgado, A., McDonald, Sk., Xu, N., & Sandler, M. (2019). A New Dataset for Amateur Vocal Percussion Analysis. Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound, 17–23. https://doi.org/10.1145/3356590.3356844

Evain, S., Lecouteux, B., Schwab, D., Contesse, A., Pinchaud, A., & Henrich Bernardoni, N. (2021). Human beatbox sound recognition using an automatic speech recognition toolkit. Biomedical Signal Processing and Control, 67, 102468. https://doi.org/10.1016/j.bspc.2021.102468

Haq, A., Nasrun, M., Setianingsih, C., & Murti, M. (2020). Speech Recognition Implementation Using MFCC and DTW Algorithm for Home Automation. Proceeding of the Electrical Engineering Computer Science and Informatics, 7, 78–85. https://doi.org/10.11591/eecsi.v7.2041

humanbeatbox.com. (2005, April 20). History of Beatbox: Old School. HUMAN BEATBOX. https://www.humanbeatbox.com/articles/history-of-beatboxing-part-2/

Icht, M. (2019). Introducing the Beatalk technique: Using beatbox sounds and rhythms to improve speech characteristics of adults with intellectual disability. International Journal of Language & Communication Disorders, 54(3), 401–416. https://doi.org/10.1111/1460-6984.12445

Jordal, I., Tamazian, A., Chourdakis, E. T., Angonin, C., Dhyani, T., askskro, Karpov, N., Sarioglu, O., BakerBunker, kvilouras, Çoban, E. B., Mirus, F., Lee, J.-Y., Choi, K., MarvinLvn, SolomidHero, & Alumäe, T. (2023). iver56/audiomentations: V0.33.0 (Version v0.33.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.8301100

Kylie, P., Michael, D., Eric, L., & Hay, K. (2011). The Nirvana Effect: Tapping Video Games To Mediate Music Learning and Interest. International Journal of Learning and Media, MIT Press. https://scholarworks.iu.edu/dspace/handle/2022/14637

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: A survey. Multimedia Tools and Applications, 80(6), 9411–9457. https://doi.org/10.1007/s11042-020-10073-7

Matveeva, L. V., Yafalian, A. F., Konovalova, S. A., & Tagiltseva*, N. G. (2019). Beatbox Subculture In The Development Of Musical Abilities Of Adolescents. European Proceedings of Social and Behavioural Sciences, Psychology of Subculture: Phenomenology and Contemporary Tendencies of Development. https://doi.org/10.15405/epsbs.2019.07.91

One Hand Clapping. (2020). [PC]. Bad Dream Games. https://handy-games.com/en/games/one-hand-clapping/

Picart, B., Brognaux, S., & Dupont, S. (2015). Analysis and automatic recognition of Human BeatBox sounds: A comparative study. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4255–4259. https://doi.org/10.1109/ICASSP.2015.7178773

Ramires, A. F. S. (2017). Automatic Transcription of Drums and Vocalised percussion. Universidade Do Porto (Portugal) ProQuest Dissertations Publishing. https://repositorio-aberto.up.pt/bitstream/10216/105309/2/200101.pdf

Ranjan, R., & Thakur, A. (2019). Analysis of Feature Extraction Techniques for Speech Recognition System. International Journal of Innovative Technology and Exploring Engineering, 8(7C2). https://www.researchgate.net/publication/343205823_Analysis_of_Feature_Extraction_Techniques_for_Speech_Recognition_System

Revd Gavin, T., & Mark, S. (2014, September 18). Standard Beatbox Notation (SBN). HUMAN BEATBOX. https://www.humanbeatbox.com/articles/standard-beatbox-notation-sbn/

Rynjah, F., Syiem, B., & L, J. S. (2022). Investigating Khasi Speech Recognition Systems using a Recurrent Neural Network-Based Language Model. International Journal of Engineering Trends and Technology, 70(7), 269–274. https://doi.org/10.14445/22315381/IJETT-V70I7P227

Saputra, F., Namyu, U. G., Vincent, Suhartono, D., & Gema, A. P. (2021). Automatic Piano Sheet Music Transcription with Machine Learning. *Journal of Computer Science*, 17(3), 178–187. https://doi.org/10.3844/jcssp.2021.178.187

Sinyor, E., McKay, C., Fiebrink, R., McEnnis, D., & Fujinaga, I. (2005). Beatbox Classification Using ACE. ISMIR 2005, 672–675. https://www.researchgate.net/publication/220722995_Beatbox_Classification_Using_ACE

Stowell, D., & Plumbley, M. D. (2010). Delayed Decision-making in Real-time Beatbox Percussion Classification. Journal of New Music Research, 39(3), 203–213. https://doi.org/10.1080/09298215.2010.512979

Tiwari, V. (2010). MFCC and its applications in speaker recognition. International Journal on Emerging Technologies, 1(1), 19–22.

Wei, S., Zou, S., Liao, F., & lang, weimin. (2020). A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. Journal of Physics: Conference Series, 1453(1), 012085. https://doi.org/10.1088/1742-6596/1453/1/012085

Wu, C.-W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Müller, M., & Lerch, A. (2018). A Review of Automatic Drum Transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(9), 1457–1483. IEEE/ACM Transactions on Audio, Speech, and Language Processing. https://doi.org/10.1109/TASLP.2018.2830113