ON DRUM PLAYING TECHNIQUE DETECTION IN POLYPHONIC MIXTURES USING NMF ACTIVATION BASED FEATURES

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT

In this paper, the problem of drum playing technique detection in polyphonic mixtures of music is addressed. We focus on the identification of 4 rudimentary techniques: strike, buzz roll, flam, and drag. The specifics and the challenges of this task are being discussed, and different sets of features are compared, including baseline spectral features, as well as various features extracted from NMF-based activation functions. We investigate the capabilities and limitations of the presented system in the case of real-world recordings and polyphonic mixtures. To design and evaluate the system, two datasets are introduced: a training dataset generated from individual drum hits, and additional annotations of the well-known ENST drum dataset minus one subset as test dataset. The results demonstrate issues with the traditionally used spectral features, and indicate the potential of using NMF activation functions for playing technique detection, however, the performance of polyphonic music still leaves room for future improvement.

1. INTRODUCTION

Automatic Music Transcription (AMT), being one of the most popular research topics in the Music Information Retrieval (MIR) community, is a process of transcribing the musical events in the audio signal into other notations such as MIDI or sheet music. In spite of being intensively studied, there are still many challenges in AMT as pointed out by Benetos et al. [1]. One of the challenges is to extract more information, such as dynamics, expressive notation and articulation, in order to produce a more complete output.

For pitched instruments, most of the work in AMT mainly focused on tasks such as melody extraction [3], chord estimation [10] and instrument recognition [8], relatively few studies try to expand the scope to playing technique and expression detection for instruments such as electric guitar [5, 16] and violin [11]. Likewise, for percussive instruments, the main focus of an AMT system has been put on recognizing the instrument types (e.g., HiHat (HH), Snare Drum (SD), Bass Drum (BD)) and their corresponding onset times [2,6,13,17,19]. Studies on retrieving the playing techniques and expressions are relatively sparse.

Since playing technique is an important layer of a musical performance for its deep connection to the timbre and subtle expressions of an instrument, an automatic system that can transcribe such techniques may provide insights to the performance and facilitate other research in MIR. In this paper, we present a system that aims to detect the drum playing techniques within the polyphonic mixtures of music. The contributions of this paper is summarized as follows: first, to the best of our knowledge, this paper is one of the earliest attempts to detect drum playing techniques in the polyphonic mixtures of music. The results may shed lights on the development towards a more complete drum transcription system. Second, a comparison between the commonly used timbre features and the NMF activation based features is presented and discussed. The experiments reveal the potential drawback of using the timbre features against the unknown testing sounds. Third, two datasets for both training and testing are introduced. These datasets may serve as the preliminary resources for the community and could be extended in the future.

The rest of the paper is structured as follows: in Section 2, related work in drum playing technique detection is introduced. The details of the proposed system and the extracted features are described in Section 3, and the evaluation process, metrics, and the experiment results are shown in Section 4. Finally, the conclusion and future directions for improvements are addressed in Section 5.

2. RELATED WORK

Percussion instrument is one of the oldest musical instruments that generates sounds through vibrations induced by strikes and other excitation [14]. While the basic gesture may be simple, the generated sounds can be complex depending on where and how the instrument is being excited. In western popular music, a drum set, which contains multiple pieces such as SD, BD, HH...etc., is one of the most commonly used percussion instruments. In general, every piece in a drum set can be excited using drum sticks. With a good control of the drum sticks, variations in timbre can be created through different excitation and gestures [15]. These gestures, sometimes referred to as rudiments, are the foundations of many drum playing techniques. As listed on the drum online learning resource [1], these rudiments can be categorized into four types:

1. **Roll Rudiments**: drum rolls created by single or multiple bounce strokes (Buzz Roll).
2. **Paradiddle Rudiments**: a mixture of alternative single and double strokes.
3. **Flam Rudiments**: drum hits with one preceding grace note.
4. **Drag Rudiments**: drum hits with two preceding grace notes created by double stroke.

Besides these rudiments, there are also other playing techniques that are commonly used to create timbral variations in a drum set, such as *Brush*, *Cross Stick*, *Rim Shot*, etc. However, these variations are usually left out in many drum transcription systems [2,6,13,17,19] in order to simplify the task.

In an early attempt to retrieve percussion gestures from the audio signal, Tindale et al. investigated the timbral variations of the snare drum sounds induced by different excitation [18]. In their study, three expert players were asked to play on different locations of the snare drums (center, halfway, edge, etc.) with different excitation (strike, rim shot, and brush), resulting a dataset with 1260 individual samples. The classification results for this dataset based on different features and classifiers were reported, and an overall accuracy around 90% was achieved. However, since the dataset is relatively small, it is difficult to generalize the results to different scenarios. Also, whether or not this method can be applied to transcribe real-world drum recordings still needs to be evaluated.

Following the same direction, Prockup et al. further explored the discrepancy between more percussion expressions with a larger dataset that covers multiple drums in a standard drum set [12]. A similar dataset was created with combinations of different drums, stick heights, stroke intensities, strike positions and articulations. Using a machine learning based approach similar to [18], various features were extracted from the samples, and a Support Vector Machine (SVM) was trained to classify the sounds. Based on the results, an accuracy of over 95% was achieved on multiple drums. However, the applicability of this approach for transcribing real-world drum recordings still needs to be tested, and the potential impact of polyphonic background music could be another concern of this approach.

Another way to retrieve more information from the drum performance is through the use of multi-modal data [9]. Hochenbaum and Kapur proposed to perform drum hand recognition using the data captured simultaneously by microphone and accelerometers. Two performers were asked to play the snare drum with four different rudiments (namely single stroke roll, double stroke open roll, single paradiddle and double paradiddle). Standard spectral and temporal features were extracted from the audio and

[1] http://vicfirth.com/40-essential-rudiments/ Last Access: 2016/3/16

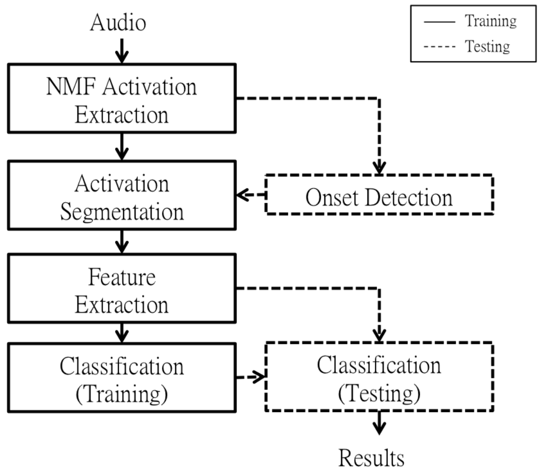

Figure 1. Block diagram of the proposed system

accelerometer data, and different classifiers were applied and compared. With Multi-Layer Perceptron (MLP), an accuracy of around 84% was achieved. However, the extra requirement of attaching the sensors on the performers' hands might alter the playing experience and thus deviate from the real playing gestures. Also, this method does not allow to analyze the existing audio recordings.

In general, the above mentioned studies mainly focused on evaluating the discriminability of the isolated samples. The evaluation on the real-world drum recordings is usually unavailable due to the lack of annotated datasets. In Table 1, different datasets for drum transcription are shown. It can be found that most of the datasets only contain annotations of playing techniques that are easily distinguishable from the normal strike (e.g., Cross Stick, Brush, Rim Shot). For playing techniques such as **Flam**, **Drag** and **Buzz Roll**, there are no datasets and annotations available. Therefore, in this paper we focus on these missing parts, augment one existing dataset with corresponding annotations, investigate differences between these techniques, and attempt to transcribe them in the real-world drum recordings under the influence of polyphonic music. The goal of this study is to enhance the existing drum transcription systems with more details on the drum performance, and facilitate research in music education, music style identification, and other related topics.

3. METHOD

3.1 System Overview

The block diagram of the proposed system is shown in Figure 1. The system consists of two stages: training and testing. During the training stage, the NMF activation functions (please see Section 3.2.1 for more details) will first be extracted from the training data. Here the training data only consists of audio clips with one-shot samples of different playing techniques. Next, a short segment around the salient peak in the activation function will be selected, and different sets of features, including the activation based features, will be computed to represent this segment. Fi-

Dataset	Annotated Techniques	Description	Total
Data in [15]	Strike, Rim Shot, Brush	1 drum (snare), 5 strike positions (from center to edge)	1264 clips
MDLib2.2 [16]	Strike, Rim Shot, Buzz Roll, Cross Stick	9 drums, 4 stick heights, 3 stroke intensities, 3 strike positions	10624 clips
IDMT-Drum [9]	Strike	3 drums (snare, bass and hihat), 3 drum kits (real, waveDrum, technoDrum)	560 clips
ENST Drum Minus One Subset [18]	Strike, Rim Shot, Brush, Cross Stick	13 drums, 3 drum kits played by 3 drummers	64 tracks

Table 1. An overview of publicly available datasets for drum transcription tasks

nally, all of the features and their corresponding labels will be used to train a classifier. For the testing, similar procedure is performed. When the entire drum recording is used as the testing data, an additional onset detection step should be taken to narrow down the area of interest. Since the focus of this paper is on playing technique detection, the onset detection step is bypassed by adopting the annotated ground truth in order to simulate the best case scenario. Once the features have been extracted from the segments, the pre-trained classifier can be used to classify the playing technique in the recordings. More details will be elaborated in the following sections.

3.2 Feature Extraction

3.2.1 Activation Functions

To detect drum playing technique in the polyphonic music, a transcription method that is relatively robust against the influence of background music would be preferable. In this paper, we applied the drum transcription scheme as described in [19] for its adaptability to polyphonic mixtures of music. The flowchart of the process is shown in Figure 2. All of the audio samples are down-mixed to mono and resampled to a sampling rate of 44.1 kHz. The Short Time Fourier Transform (STFT) of the input signal is computed with a block size of 512 and a hop size of 128 for better temporal resolution, and the hann window is applied on each block.

The magnitude spectrogram is decomposed using the partially fixed NMF with a harmonic rank $r_h = 50$ and a pre-trained drum dictionary matrix built from the ENST drum dataset [7], and the corresponding activation functions of each drum can be obtained. For each individual drum, the activation function $h_i(n)$ is a 1 by n vector, in which n is the block index and $i = \{HH, SD, BD\}$ indicates the type of drum. The resulting $h_i(n)$ is scaled to a range between 0 and 1 and smoothed using a median filter with an order of $p = 5$ samples. Since a template in the dictionary is intended to capture the activity of the same type of drum, the drum sounds with slightly different timbres could still result in similar $h_i(n)$. Therefore, the extracted activation function $h_i(n)$ can be considered as a timbre invariant transformation and is desirable for detecting the underlying techniques. These activation functions can be used as features directly or as the intermediate representation for deriving more features.

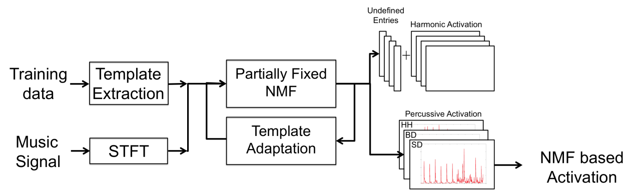

Figure 2. Flowchart of the activation extraction process in [19]

3.2.2 Activation Derived Features

Once the activation functions $h_i(n)$ have been extracted from the audio data, various features can be derived for further classification. The steps can be summarized as follows: first, for every given onset at index n_o, a 400 ms segment centered around $h_i(n_o)$ will be selected. Next, the segment will be slightly adjusted to ensure the maximum value is at the center. Based on the adjusted segment, the distribution features, Inter-Onset Interval (IOI) features, and peak features are computed for pattern recognition.

The distribution features are similar to the commonly used spectral features, which provide the general description of the pattern. The IOI features of the activation function are designed to capture the general temporal consistency of the pattern, whereas the peak features are designed to describe the details of the pattern. To compute the peak features, first we find the local maximums and sort them in descending order, then we calculate the ratio and index difference between the side peak and the main (largest) peak as features. Since we are looking for specific patterns in the activation functions to identify the playing techniques, the cumulative costs from the Dynamic Time Warping (DTW) are also used as features. To compute the DTW features, a median activation template of each playing technique is trained from the training data, and the cumulative cost of every DTW template for the given segment can be calculated. The examples of the extracted DTW templates for each technique are shown in Figure 3. Finally, the resulting feature vector has a dimension $d = 19$. The features are listed as follows:

1. **Distribution features,** $d = 5$: Spread, Skew, Crest, Centroid, Flatness

2. **IOI features,** $d = 2$: IOI mean, IOI standard deviation

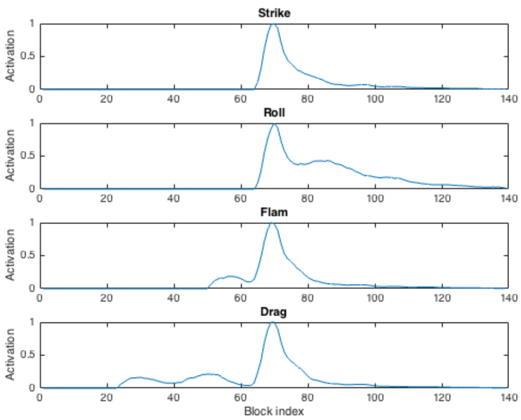

Figure 3. Examples of the extracted activation functions of (Top to bottom): Strike, Buzz Roll, Flam, Drag

3. **Peak features,** $d = 8$: side peak to main peak ratio α_i, side peak to main peak signed block index difference Δb_i $i = \{1, 2, 3, 4\}$

4. **DTW features,** $d = 4$: cumulative cost of each DTW templates

3.2.3 *Timbre Features*

To compare the effectiveness of the activation based features, a set of the commonly used timbre features is also extracted. The extraction process is similar to Section 3.2.2, however, instead of using activation functions, the time-domain waveform of a given segment is used to derive the features. The features are listed as follows:

1. **Spectral features,** $d = 3$: Centroid, Rolloff, Flux

2. **Temporal features,** $d = 1$: Zero crossing rate

3. **MFCCs,** $d = 13$: the first 13 MFCC coefficients

The features are computed block by block using the same parameters as described in Section 3.2.1. The resulting feature vectors are further aggregated into one single vector by computing the mean and standard deviation of all the blocks. The final feature vector has a dimension $d = 34$.

3.3 Dataset

3.3.1 *Training Dataset*

In this paper, we focus on four different playing techniques (Strike, Flam, Drag, Buzz Roll) played on the snare drum. As can be seen in Table 1, only Strike and Buzz Roll can be found in some of these datasets. To further detect Flam and Drag, we constructed a dataset by augmenting the existing MDLib 2.2 [12] with synthesized audio clips. Since both Flam and Drag consist of preceding grace notes with different velocity and timing, they could be modeled with a limited set of parameters as shown in Figure 4. The blue

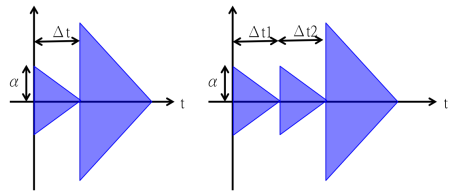

Figure 4. Illustration of the parametric forms of (Left) Flam and (Right) Drag

Techniques	Description	Total (#clips)
Strike	Snare excerpts from MDLib 2.2 [16]	576
Buzz Roll	Snare excerpts from MDLib 2.2 [16]	576
Flam	144 mono snare excerpts α = {0.1:0.1:0.7} Δt = {30:10:60} (ms)	4032
Drag	144 mono snare excerpts α = {0.15:0.1:0.55} Δt_1 = {50:10:70} (ms) Δt_2 = {45:10:75} (ms)	8640

Table 2. An overview of the constructed dataset

triangles in the figure represents the basic waveform excited by normal strikes, and the Δt is the time difference between the preceding grace note and the strong note. All the waveforms have been normalized to a maximum amplitude of -1 to 1, and the α is the amplitude ratio between the grace note and the strong note.

To synthesize the audio signals with settings for Δt and α, we annotated the demo videos from Vic Firth's online lessons for both Flam [2] and Drag [3]. The final parameter settings and the details of the constructed dataset are shown in Table 2. The parameters are based on the mean and standard deviation estimated from the videos. The resulting data contains all possible combinations of the parameters with the 144 mono snare Strike in the MDLib 2.2. However, to ensure the classifier is trained with uniformly distributed classes, only 576 randomly selected clips are used for Flam and Drag during the training. All of generated training data is available upon request.

3.3.2 *Testing Dataset*

To evaluate the system for detecting the playing techniques in the polyphonic mixtures of music, the tracks from the ENST drum dataset minus one subset [7] have been annotated. The ENST drum dataset contains various drum recordings from 3 drummers with 3 different drum kits. The minus one subset, specifically, consists of 64 tracks of drum recordings with individual channel, mix, and accompaniments available. Since the playing technique is related to the playing style of the drummer, only 30 out of 64 tracks contain such techniques on snare drum. These techniques are annotated using the snare channel of the recordings, and each technique is labeled with the starting time, duration, and the technique index. As a result, a total number of 182 events (Roll: 109, Flam: 26, Drag:

[2] http://vicfirth.com/20-flam/ Last Access: 2016/03/16
[3] http://vicfirth.com/31-drag/ Last Access: 2016/03/16

47) have been annotated, and each event has a length of approximately 250 to 400 ms. The annotations are available online. [4]

4. EVALUATION

4.1 Metrics

For evaluating the classification accuracy on the training data, we applied the 10-fold cross validation and compute the overall accuracy. For evaluating the accuracy on the testing data, however, we calculated the *micro-averaged accuracy* and the *macro-averaged accuracy* [20] to account for the unevenly distributed and sparse classes. The metrics are defined in the following equations:

$$micro\ averaged = \frac{\sum_{k=1}^{K} C_k}{\sum_{k=1}^{K} N_k} \tag{1}$$

$$macro\ averaged = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{C_k}{N_k}\right) \tag{2}$$

In which K is the total number of classes, N_k is the total number of samples in class k, and C_k is the total number of correct samples in class k. These two metrics have different interpretations: in micro-averaged accuracy, each sample is weighted equally, whereas in macro-averaged accuracy, each class is weighted equally. These two metrics give a better overview of the performance by emphasizing the minority classes.

4.2 Experiment Results

In this section, results from three experiments are presented. The first experiment is the cross-validation results from the training data, the second experiment is the detection results from the testing data with the annotation informed segmentation, and the third experiment is the detection results from the testing data without the annotation informed segmentation. Different feature sets as described in Section 3.2, namely Activation Functions (AF), Activation Derived Features (ADF), and Timbre Features (TF), are tested using a multi-class C-SVM with Radial Basis Function (RBF) kernel and default parametrization. For the implementation, we used *libsvm* [4] in Matlab. All of the features are scaled to a range between 0 and 1 using the standard min-max scaling approach before feeding into the SVM functions.

4.2.1 Experiment 1: Cross-Validation of Training Data

In Experiment 1, the 10-fold cross validation on the training data using different sets of features are performed, and the results are shown in Figure 5. This experiment is similar to the approaches described in the previous work [12, 18], where it provides evidence to the discriminability of the feature representations. Since the training data contains 576 samples for all classes, the micro-averaged and marco-averaged accuracy are the same.

[4] http://dummy link

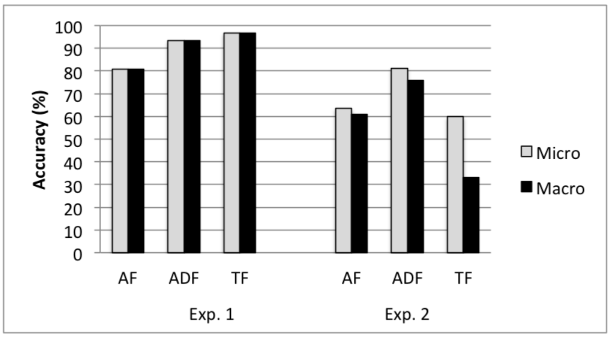

Figure 5. Results of experiment 1 (left) and experiment 2 (right)

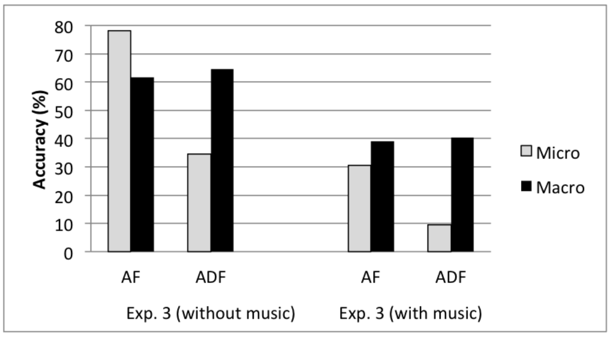

Figure 6. Results of experiment 3 without background music (left) and with background music (right)

4.2.2 Experiment 2: Annotation-Informed Testing

In Experiment 2, the same sets of features are extracted from the testing data for evaluation. Since the testing data is a completely different dataset with the real-world drum recordings, a verification of the feasibility of using the synthetic training data as well as the proposed feature representations is necessary. For this purpose, we simulate the best case scenario by using the snare channel as the input with an annotation-informed segmentation process. The resulting 182 segments are then classified using the trained SVM models from Experiment 1. This experiment serves as a sanity check to the presented scheme, and the results are shown in Figure 5.

4.2.3 Experiment 3: Real-World Testing

In Experiment 3, a more realistic setup is applied, where each onset is examined and classified without any prior knowledge about the segmentation. In this case, a fixed region around every onset is segmented and classified. As a result, a total number of 2943 onsets (including the previous mentioned 182 playing technique events and 2761 strikes) are evaluated. Since the timbre features do not show promising results in Experiment 2, they are excluded from this experiment. To study the influence of the background music, both the recordings of the snare channel and the complete polyphonic mixtures are tested. The results are shown in Figure 6.

	Strike	Roll	Flam	Drag
Strike	28.9	38.8	5.7	26.6
Roll	8.3	66.1	11.9	13.8
Flam	3.8	53.8	19.2	23.1
Drag	46.8	6.4	4.3	42.6

Table 3. Confusion matrix of {Exp.3 with music, AF}

	Strike	Roll	Flam	Drag
Strike	5.8	62.9	3.8	27.6
Roll	5.5	74.3	3.7	16.5
Flam	0.0	61.5	11.5	26.9
Drag	2.1	8.5	19.1	70.2

Table 4. Confusion matrix of {Exp.3 with music, ADF}

4.3 Discussion

Based on the experiment results, the following observations can be made:

First, as can be seen in Figure 5, the timbre features achieve the highest cross-validation accuracy in Experiment 1, which shows its effectiveness of differentiating different groups of sounds. However, when these features are applied to classify a completely different dataset, its inability of recognizing the same pattern played with different drum sounds is revealed. As a result, the timbre features achieve the lowest macro-averaged accuracy in Experiment 2, and the micro-averaged accuracy only reflects the accuracy by always predicting the majority class. The activation functions and activation derived features, on the opposite, are relatively consistent between the micro and macro-averaged accuracy, which indicates a better performance for detecting the proposed playing techniques in the unseen dataset. This observation echos the results from the related work, which demonstrate the usefulness of timbre features for distinguishing the different sounds, but it also shows that timbre features might not be directly applicable to detecting the playing techniques in the unknown recordings.

Second, compared with the activation functions, the activation derived features are more sensitive to different playing techniques, whereas the activation functions are more sensitive to strikes. Among all the experiment results, the activation derived features tend to achieve a higher macro-averaged accuracy than the activation functions. These results show that the activation derived features are more capable of detecting the playing techniques. This tendency can also be shown in the confusion matrix in Figure 3 and Figure 4, in which the activation derived features perform better than the activation functions in Roll and Drag, and slightly worse in Flam. In Experiment 3, it is shown that the activation functions generally achieve higher micro-averaged accuracy than the activation derived features. Since we know the distribution of the classes is skewed towards Strike in the testing data, we found that the the micro-averaged accuracy of the activation functions is largely increased by a higher rate of detecting strikes.

Third, according to Figure 3 and Figure 4, Strike and Flam can be easily confused with Roll for both features in the context of polyphonic mixtures of music. One possible explanation is that, whenever the signal is not properly segmented, the activation function will contain unfinished activities from the previous or the next onset, which might add tails to the original activation and make it resemble a Roll. As for Drag, since the strong features is the preceding grace notes, it is relatively easy to distinguish from Roll for both features.

Fourth, for both activation functions and activation derived features, the detection performance drops drastically in Experiment 3 with the presence of background music. The reason could be that with the background music, the extracted activation function becomes noisier due to the imperfect decomposition. Since the classification models are trained on the clean signals, they might be susceptible to these disturbance. As a results, the classifier might be tricked into classifying Strike as other playing techniques, which decreases the micro-averaged accuracy.

Additionally, the proposed method does not take into account the onset detection at this moment. By adding the onset detection process to the loop, the detection accuracy could be further reduced, which increases the difficulty of direct application of the method in the polyphonic mixtures. To build a system towards retrieving the playing techniques reliably from the polyphonic mixtures of music, further improvements are necessary.

5. CONCLUSION

In this paper, a system working towards drum playing technique detection in the polyphonic mixtures of music has been presented. To achieve this goal, two datasets have been generated for training and testing purposes. The experiment results indicate that the current method is able to detect the playing techniques from the real-world drum recordings when the signal is relatively clean. However, the system suffers from the presence of background music and has to be improved in order to detect the playing techniques in the polyphonic mixtures of music with higher accuracy.

The possible directions for the future work are: first, test different source separation algorithms as a pre-processing step in order to get a cleaner representation. When comparing the results in Experiment 3, it is shown that a cleaner representation would improve both the micro and macro-averaged accuracy by over 20%. Therefore, a good source separation method to isolate the snare drum sound could be beneficial. Common techniques such as HPSS and NMF for source separation could be further investigated.

Second, since the results in Experiment 3 implies that the system is susceptible to the disturbance from background music, a classification model trained on the slightly noisier data might increase the robustness against the presence of unwanted sounds. The influence of adding different levels of random noise while training could be tested.

Third, the current dataset only offers a limited number of samples for evaluating playing detection in the polyphonic mixtures of music. Due to the sparse nature of these playing techniques, their presences in the existing datasets are rare and therefore difficult to collect and annotate. However, to arrive at a statistically meaningful conclusion, more data would be necessary.

Last but not least, different state-of-the-art classification methods, such as deep neural network, could also be tested in searching for a better solution.

6. REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, jul 2013.

[2] Emmanouil Benetos, Sebastian Ewert, and Tillman Weyde. Automatic transcription of pitched and unpitched sounds from polyphonic music. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2014.

[3] Rachel M. Bittner, Justin Salamon, Slim Essid, and Juan P. Bello. Melody extraction by contour classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 500–506, 2015.

[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[5] Yuan-ping Chen, Li Su, and Yi-hsuan Yang. Electric Guitar Playing Technique Detection in Real-World Recordings Based on F0 Sequence Pattern Recognition. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 708–714, 2015.

[6] Christian Dittmar and Daniel Gärtner. Real-time Transcription and Separation of Drum Recording Based on NMF Decomposition. In *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, pages 1–8, 2014.

[7] Olivier Gillet and Gaël Richard. Enst-drums: an extensive audio-visual database for drum signals processing. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2006.

[8] Philippe Hamel, Sean Wood, and Douglas Eck. Automatic Identification of Instrument Classes in Polyphonic and Poly-Instrument Audio. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 399–404, 2009.

[9] Jordan Hochenbaum and A Kapur. Drum Stroke Computing: Multimodal Signal Processing for Drum Stroke Identification and Performance Metrics. *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2011.

[10] Eric J Humphrey and Juan P Bello. Four timely insights on automatic chord estimation. *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2015.

[11] Pei-Ching Li, Li Su, Yi-hsuan Yang, and Alvin W. Y. Su. Analysis of Expressive Musical Terms in Violin using Score-Informed and Expression-Based Audio Features. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2015.

[12] Matthew Prockup, Erik M. Schmidt, Jeffrey Scott, and Youngmoo E. Kim. Toward Understanding Expressive Percussion Through Content Based Analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2013.

[13] Axel Roebel, Jordi Pons, Marco Liuni, and Mathieu Lagrange. On Automatic Drum Transcription Using Non-Negative Matrix Deconvolution and Itakura Saito Divergence. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[14] Thomas D Rossing. Acoustics of percussion instruments: Recent progress. *Acoustical Science and Technology*, 22(3):177–188, 2001.

[15] George Lawrence Stone. *Stick Control: for the Snare Drummer*. Alfred Music, 2009.

[16] Li Su, Li-Fan Yu, and Yi-Hsuan Yang. Sparse ceptral and phase codes for guitar playing technique classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, number Ismir, pages 9–14, 2014.

[17] Lucas Thompson, Matthias Mauch, and Simon Dixon. Drum Transcription via Classification of Bar-Level Rhythmic Patterns. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2014.

[18] Adam R. Tindale, Ajay Kapur, George Tzanetakis, and Ichiro Fujinaga. Retrieval of percussion gestures using timbre classification techniques. In *in Proceedings of the International Conference on Music Information Retrieval*, pages 541–544, 2004.

[19] Chih-Wei Wu and Alexander Lerch. Drum Transcription Using Partially Fixed Non-Negative Matrix Factorization with Template Adaptation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 257–263, 2015.

[20] Yiming Yang. An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval*, 1:69–90, 1999.