

ON DRUM PLAYING TECHNIQUE DETECTION IN POLYPHONIC MIXTURES

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), one of the most popular research topics in the Music Information Retrieval (MIR) community, is the process of transcribing the musical events in the audio signal into a notation such as MIDI or sheet music. In spite of being intensively studied, there still remain many unsolved problems and challenges in AMT [?]. One of the challenges is the extraction of additional information, such as dynamics, expressive notation and articulation, in order to produce a more complete description of the music performance.

For pitched instruments, most of the work in AMT mainly focuses on tasks such as melody extraction [?], chord estimation [?], and instrument recognition [?]. Few studies try to expand the scope to playing technique and expression detection for instruments such as electric guitar [?, ?] and violin [?]. Similarly, the main focus of AMT systems for percussive instruments has been put on recognizing the instrument types (e.g., HiHat (HH), Snare

Drum (SD), Bass Drum (BD)) and their corresponding onset times [?, ?, ?, ?, ?]. 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 transcribes such techniques may provide insights into the performance and facilitate other research in MIR. In this paper, we present a system that aims to detect the drum playing techniques within polyphonic mixtures of music. The contributions of this paper can be summarized as follows: first, to the best of our knowledge, this is the first study to investigate the automatic detection drum playing techniques in polyphonic mixtures of music. The results may support the future development of a complete drum transcription system. Second, a comparison between the commonly used timbre features and features based on activation functions of a Non-Negative Matrix Factorization (NMF) system are presented and discussed. The results reveal problems with using common timbre features. Third, two datasets for training and testing are introduced. The release of these datasets is intended to encourage future research in this field. The data may also be seen as a core compilation to be extended in the future.

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

2. RELATED WORK

Terminology: is gestures the right word, later you seem to use gestures as a synonym to playing technique as well? Is piece the right word for an instrument in a drum set? Percussive instruments, generating sounds through vibrations induced by strikes and other excitations, belong to the oldest musical instruments [?]. While the basic gesture is generally 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 is excited using drum sticks. With good control of the drum sticks, variations in timbre can



be created through different gestures and excitation methods and gestures [?]. These gestures, sometimes referred to as rudiments, are the foundations of many drum playing techniques. These rudiments can be categorized into four types:¹

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.

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. Most drum transcription systems, however, focus on single strikes instead of these playing techniques [?, ?, ?, ?, ?].

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 excitations [?]. Three expert players were asked to play on different locations on the snare drums (center, halfway, edge, etc.) with different excitation (strike, rim shot, and brush), resulting in a dataset with 1260 individual samples. The classification results for this dataset based on different features and classifiers **elaborate on features and classifiers?** were reported, and an overall accuracy of around 90% was achieved. Since the dataset is relatively small, however, it is difficult to generalize the results to different scenarios. The method was not evaluated with real-world drum recordings. **maybe combine summary with following paper?**

Following the same direction, Prockup et al. further explored the discrepancy between more percussion expressions **now is it gestures or technique or expression?** with a larger dataset that covers multiple drums of a standard drum set [?]. A dataset was created with combinations of different drums, stick heights, stroke intensities, strike positions and articulations. Using a machine learning based approach similar to [?], various features were extracted from the samples, and a Support Vector Machine (SVM) was trained to classify the sounds. An accuracy of over 95% was reported on multiple drums **more specifics?** How many classes, what features. 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 [?]. Hochenbaum and Kapur investigated the inclusion of drum hand recognition in the data by capturing microphone and accelerometer data simultaneously. Two performers were asked to play the snare drum with four different rudiments (namely single stroke roll, double stroke

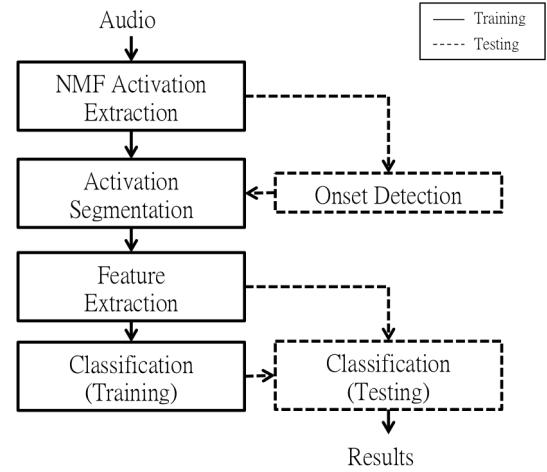


Figure 1. Block diagram of the proposed system

open roll, single paradiddle and double paradiddle). Standard spectral and temporal features were extracted from the audio and accelerometer data, and different classifiers were applied and compared. With a Multi-Layer Perceptron (MLP), an accuracy of around 84% was achieved. again, more info on features and classes? It cannot be ruled out that the extra requirement of attaching the sensors on the performers' hands might alter the playing experience and thus deviate from the real playing gestures. Furthermore, this method does not allow the analysis of existing audio recordings.

In general, the above mentioned studies mainly focused on evaluating the discriminability of isolated samples. The evaluation on real-world drum recordings, i.e. recordings of a drummer continuously playing, is usually unavailable due to the lack of annotated datasets. In Table 1, different datasets for drum transcription are presented. 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** why bold? — is this the introduction of your classes. Then do it more systematically in Method, there are no datasets and annotations available.

the following is not related work. Find a better place 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 from 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 test-

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

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

ing. During the training stage, NMF already introduced? activation functions (see Sect. 3.2.1) 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, features will be extracted from a short segment around the salient peak in the activation function (see Sect. XXX). Finally, all of the features and their corresponding labels will be used to train a classifier. introduce the classes here For the testing, a similar procedure is performed. When a longer drum recording is used as the testing data, an additional onset detection step is taken to narrow down the area of interest. the following sentence maybe later — not really system overview 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 given 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 [?] 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 [?], 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

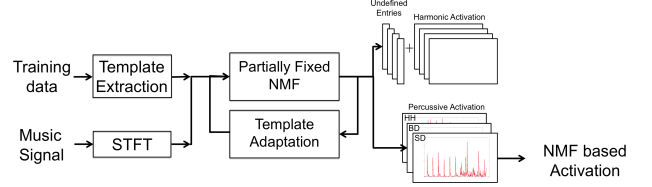


Figure 2. Flowchart of the activation extraction process in [?]

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.

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 result-

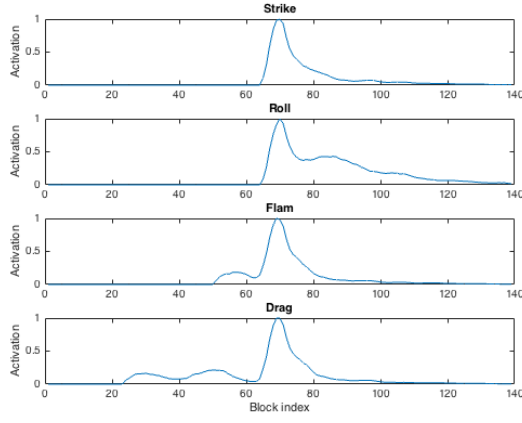


Figure 3. Examples of the extracted activation functions of (Top to bottom): Strike, Buzz Roll, Flam, Drag

ing 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
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 Sect. 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 Sect. 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 $d = 34$.

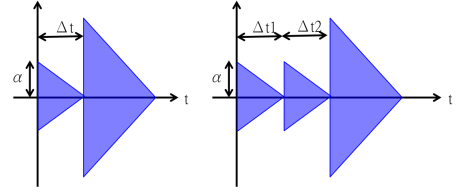


Figure 4. Illustration of the parametric forms of (Left) Flam and (Right) Drag

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 $\alpha = \{0.1:0.1:0.7\}$ $\Delta t = \{30:10:60\}$ (ms)	4032
Drag	144 mono snare excerpts $\alpha = \{0.15:0.1:0.55\}$ $\Delta t_1 = \{50:10:70\}$ (ms) $\Delta t_2 = \{45:10:75\}$ (ms)	8640

Table 2. An overview of the constructed dataset

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 [?] 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 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 Flam² and Drag³. 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

² <http://vicfirth.com/20-flam/> Last Access: 2016/03/16

³ <http://vicfirth.com/31-drag/> Last Access: 2016/03/16

ENST drum dataset minus one subset [?] 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: 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* [?] to account for the unevenly distributed and sparse classes. The metrics are defined in the following equations:

$$\text{micro averaged} = \frac{\sum_{k=1}^K C_k}{\sum_{k=1}^K N_k} \quad (1)$$

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

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 Sect. 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* [?] 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 <http://dummy link>

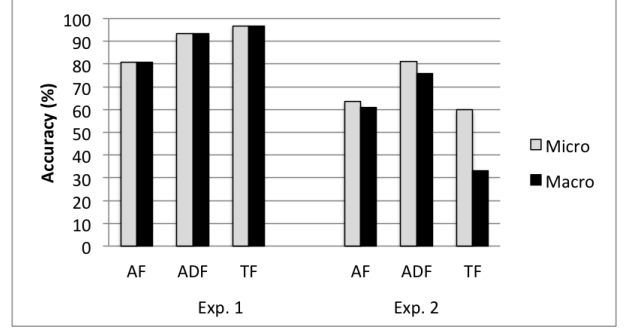


Figure 5. Results of experiment 1 (left) and experiment 2 (right)

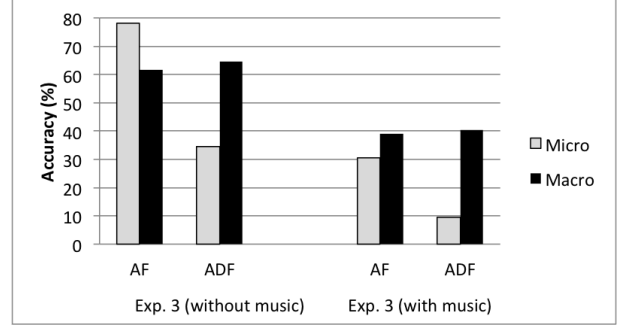


Figure 6. Results of experiment 3 without background music (left) and with background music (right)

4.2.1 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 [?, ?], 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 macro-averaged accuracy are the same.

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.

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 pre-

	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}

ceding 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