2015-07

Automatic classification of drum sounds with indefinite pitch

International Joint Conference on Neural Network, 2015, Killarney.
http://www.producao.usp.br/handle/BDPI/49424

Automatic Classification of Drum Sounds with Indefinite Pitch

Vinícius M. A. Souza
Gustavo E. A. P. A. Batista
Institute of Mathematics and Computer Science
University of São Paulo, Brazil
{vsouza, gbatista}@icmc.usp.br

Nilson E. Souza-Filho
Department of Acoustic Engineering
Federal University of Santa Maria, Brazil
nilson.evilasio@eac.ufsm.br

***Abstract*—Automatic classification of musical instruments is an important task for music transcription as well as for professionals such as audio designers, engineers and musicians. Unfortunately, only a limited amount of effort has been conducted to automatically classify percussion instrument in the last years. The studies that deal with percussion sounds are usually restricted to distinguish among the instruments in the drum kit such as toms *vs.* snare drum *vs.* bass drum *vs.* cymbals. In this paper, we are interested in a more challenging task of discriminating sounds produced by the same percussion instrument. Specifically, sounds from different drums cymbals types. Cymbals are known to have indefinite pitch, nonlinear and chaotic behavior. We also identify how the sound of a specific cymbal was produced (e.g., roll or choke movements performed by a drummer). We achieve an accuracy of 96.59% for cymbal type classification and 91.54% in a classification problem with 12 classes which represent the cymbal type and the manner or region that the cymbals are struck. Both results were obtained with Support Vector Machine algorithm using the Line Spectral Frequencies as audio descriptor. We believe that our results can be useful for a more detailed automatic drum transcription and for other related applications as well for audio professionals.**

I. Introduction

In general, musicians have great interest in reading musical transcriptions concerning their instrument. However, in most cases the transcriptions are performed manually by other musicians. This task is time consuming and requires the availability of a skilled person to transcribe a particular music. Thus, the automatic identification and classification of particularities of musical instruments is a important step in direction to automatic transcription of these instruments in polyphonic music.

Automatic classification of instruments is also interesting for audio designers. The amount of audio samples available on internet that can be used for music creation grows every day. However, most of these data are unlabeled or have insufficient information. Consequently, the user is frequently obligated to listen many audio files for selecting only a small subset of interest. Thus, the large amount of data makes this task costly and tiresome. Consequently, the automatic classification can be very useful for these professionals.

In addition to these examples, we can also mention other applications that may benefit from automatic classification of instrument sounds such as those related to audio content analysis or for professionals such as audio engineers and musicians in the mixing process during recording sessions.

Many research papers in Machine Learning and Signal Processing literature focus in the classification of string or wind harmonic instruments and only a limited effort has been conducted to classify percussion instruments (an interesting review can be found in [1]). The main difference between percussion and another instruments is the fact that the percussion produces indefinite pitch or unpitched sounds. Pitch is a perceptual property that allows the ordering of sounds on a frequency-related scale [2]. Although some pieces of the drum such as toms, bass drum and snare drum can be tuned by the player, this tuning does not relate to producing a perceived pitch achieved by other instruments. Thus, the classification task is more challenging when this property is undefined.

The studies that deal with percussion sounds are more interested on distinguishing different instruments in the drum kit such as bass drum, snare drum, hi-hat, toms and cymbals. Some examples can be seen in [3], [4], [5], [6]. Differently from these works, we are interested in discriminating the sounds produced by the same percussion idiophone instrument. More specifically, sounds produced by different cymbals types such as China, Crash, Hi-hat, Ride and Splash. Due to the high perceptual similarity of the sounds produced by the same instrument, the investigated task in this work can be considered a more difficult problem than to distinguish different instruments.

In this sense, the most related work to ours is presented by [7]. The authors proposed the use of spectral features from non-negative matrix factorization to train an 1-Nearest Neighbor algorithm to classify specific combinations of cymbals (for instance, Splash *vs.* China or Splash *vs.* China *vs.* Crash) with a very limited amount of data to train and test the classifier.

In this paper, we propose a more challenging investigation with a two-level classification of cymbal sounds. In the first level we classify the cymbal type and in the second level we identify how the sound of this cymbal was produced. To the best of our knowledge, no other work in literature has investigated the classification of percussion sounds with this specificity. Our main goal is to aid an important step on the build of systems for detailed drum transcription from polyphonic music. However, it is important to note that this work also can help audio professionals and other audio applications dependent of automatic classification of percussion sounds.

We investigate different signal processing methods combined with supervised machine learning algorithms. Moreover,

978-1-4799-1959-8/15/$31.00 ©2015 IEEE

we built a dataset with more than one thousands real cymbal sounds from a large variety of material and size and make it publicly available for other interested researchers.

Our results show that we can distinguish five different cymbals types with 96.59% accuracy. We also show that our approach can achieve 91.54% accuracy in the classification of 12 classes representing cymbal types and the manner or region that the cymbals are struck. Both results were achieved with Support Vector Machine algorithm using Line Spectral Frequencies as audio descriptor.

The main contributions of this paper are in two directions. First, we make available a dataset of real samples of drum cymbals to the community interested in unpitched musical sounds. Second, we show that the Line Spectral Frequencies (LSF) are good descriptors for these data. It is important to note that Mel-Frequencies Cepstrum Coefficients (MFCC) are considered many times state-of-the-art for a related drum transcription task [8], [9], [10]. We believe that the good performance of LSF can also be achieved in other similar data or classification tasks. Thus, we recommend whenever possible, the comparative performance evaluation between MFCC and LSF. Furthermore, we show that simple features from temporal domain can significantly improve the results achieved by LSF.

The remaining of this paper is organized as follows. In Section II we briefly discuss the main characteristics of unpitched drum sounds. In Section III we discuss some works related to classification of drum sounds. In Section IV we present the collected dataset concerning cymbals sounds and the signal processing methods used for the features extraction step. In Section V we present our experimental evaluation. Finally, in Section VI we present our conclusions and future directions for this work.

II. Unpitched Drum Sounds

Psychoacoustics studies the relationship between auditory system and physical characteristics of the sound [11]. The auditory sensations are determined by characteristics as frequency, amplitude and temporal features. For music signals, it is also considered characteristics such as *loudness, timbre* and *pitch.* Specifically, *pitch* is a perceptual property that allowing distinguish a bass sound from a treble sound. Thus, the perception of pitch is what guarantees that two distinct sounds with similar intensities can be distinguished as heterogeneous by the listener.

String or wind harmonic instruments generates sounds with well defined pitch. However, percussion instruments such as drums and specifically cymbals, produce sounds with undefined pitch. Thus, the listener finds impossible or relatively difficult to identify this perceptual characteristic in these sounds. This difficulty is extended for systems that seek automatically identify and classify the sounds generated by these instruments.

Currently, it is probable that the most popular percussion instrument are the drums. Drums play an important role in contemporary music in nearly all musical cultures. In musical genres such as *Rock, Pop* and *Jazz,* the drum is an essential instrument responsible for the rhythmic structure of the music. The pieces of a drum can be divided in *membranophones* and *idiophones.* The first one has a membrane (skin) stretched over an opening cavity (e.g., bass drum or tom-toms). The second category are rigid bodies that vibrating as a whole, called cymbals. This paper have a special interest in cymbal sounds due to their peculiarities as discussed in the follow.

Cymbals are thin, axially symmetric, isotropic (uniform in all directions) metal plates. When struck with a drumstick, waves radiate away from the excitation area at a velocity inversely proportional to the dimensions of the initial flexural indentation of the surface made by the drumstick. Both the initial pulse and the reflections off the edges of the cymbal travel up the transitional section to the bell. They are dispersed across the entire body of the cymbal, causing it to vibrate [12].

Cymbals are made from some variety of copper alloy, not only due their malleability, but because copper has desirable sonic properties. The most common copper alloys used in cymbals are from bronze, which are alloys of copper and tin with trace amounts of other metals such as silver. For example, the bell bronze or B20 bronze (80% copper, 20% tin) and B8 bronze (92% copper, 8% tin) are the most popular copper alloys. However, the cymbals' manufacturers have experienced different tin-to-copper ratios in their products along the years.

Sometimes cymbals are denoted as *nonlinear percussion instruments*, which emphasizes the fact that nonlinearity is essential in the production of sound by such sources [13]. There are evidences that the vibrations of cymbals exhibit chaotic behavior [14]. When struck hard with a stick, the cymbal oscillates chaotically, the energy spread over multiple vibrational modes. Even when struck lightly the pitch is not very clear as the cymbals are inharmonic, with many different and unrelated resonant frequencies. Because of its nonlinear behavior, the cymbal is a difficult, but not impossible, musical instrument to model [14].

III. Related Work

There are many research concerning automatic classification of instruments' sounds, some examples are [1], [15], [16], [17], [18]. Most of these works are focused in instruments with well defined pitch such as string instruments (guitar, bass, piano, cello, violin, harp, etc) or wind harmonic instruments (horns, trumpets, trombones, oboes, clarinets, etc).

On the other hand, even with the importance of percussion instruments in the contemporary music, there are a minor interest for these instruments in research fields of computer science as machine learning. There are a wide variety of percussion instruments, some examples that can be studied are vibraphone, xylophone, marimba, cabasa, cajon, bongo, conga. In this paper, we focus on the classification of drum sounds for polyphonic audio signals and in this section we present some related works in this direction.

In [3] the authors have presented a comparative evaluation of two feature selection methods (Correlation-based Feature Selection – CFS and ReliefF) and five classification techniques (Canonical Discriminant Analysis – CDA, K*, C4.5, PART and KNN) on 634 drum sounds. The sound descriptors considered are from different categories as attack-related and decay-related descriptors, relative energies for selected bands and

Mel-Frequency Cepstral Coefficients. The authors evaluated three levels of taxonomic classification: *i)* membranes *vs.* plates; *ii)* kick *vs.* snare *vs.* hi-hat *vs.* toms *vs.* cymbals; *iii)* kick *vs.* snare *vs.* ride *vs.* crash *vs.* hi-hat open *vs.* hi-hat closed *vs.* high tom *vs.* medium tom *vs.* low tom. They achieved accuracies of 99%, 97% and 90%, for these levels respectively. The authors consider only two cymbal types (ride and crash) plus the hi-hat (considered as a different class from cymbals). As a the paper have your main focus in feature selection methods, a more detailed discussion about the interclass error rates are not presented.

A system called *AdaMast* that recognize drum sounds in polyphonic audio signals is presented in [5]. The system detects the onset times and identifies three drum instruments: bass drum, snare drum, and hi-hat cymbals. The system uses a template-adaption and a template-matching method that uses power spectrograms of drum sounds as template. The method calculates the distance between a template and each spectrogram segment extracted from a song spectrogram using the Goto's distance measure. In their experimental evaluation, the authors report that the Average F-Measures were 82.92%, 58.29%, and 46.25% in recognizing bass drum sounds, snare drum sounds, and hi-hat cymbals sounds, respectively. Although the *AdaMast* considers a audio signal with different instruments (not only percussion instruments), we can note that for cymbals it is considered only the hi-hat. Even so, we can see that the hi-hat results are below to the other instruments.

Different from audio recognition works in general that use spectral features, the authors in [6] have presented a biologically inspired approach to classify drum sounds by *i)* learning sparse atomic functions on unlabeled data and *ii)* supervised learning of the classes using features derived from a temporal approximation of the signal. In *i)* they used gammatone filterbanks as atomic functions. The authors evaluated their method in a dataset with a few more than 200 audio samples that have signals from bass drum, snare drum, open hi-hat and closed hi-hat. The classification task was performed by a Support Vector Machine algorithm using a radial basis function (RBF) kernel. The classification accuracies were over 95%.

Although the above discussed works deal with cymbal sounds, we can note that they have a limited coverage. In general, only the hi-hat cymbals are focus of the classification task while the same task with membranophones' instruments (snare drums, bass drums and toms) is well established. To fill this gap, we are interested in a more deeper analysis of cymbal sounds in this paper. In this sense, the most similar work to ours found in literature is presented by [7]. The authors proposed the use of spectral features from non-negative matrix factorization to train an 1-Nearest Neighbor algorithm to classify specific combinations of cymbals (for instance, Splash *vs.* China or Splash *vs.* China *vs.* Crash). They used between 11 and 16 audio samples for each considered cymbal combination, a very limited amount of data to evaluate a supervised machine learning algorithm. The best classification rates achieved were 95% when distinguishing between two cymbals and 86% when distinguishing between three cymbals.

IV. Cymbal Dataset

In this section we describe the collected dataset with real cymbal audio samples and the feature sets used in our experimental evaluation.

A. Dataset Description

In this study we collected 1,052 samples from the traditional Swiss manufacturer and designer of cymbals *Paiste*. The sound samples are freely available in their website[1]. We collected data from 18 different cymbal series produced by the company. Each cymbal series is specified by a set of parameters such as construction material (brass or bronze alloys), texture, curvature, thickness, and diameter (from 8 to 24 inches).

Initially, we sorted the samples in five main categories according to the cymbal type: *i)* China; *ii)* Crash; *iii)* Hi-hat; *iv)* Ride and *v)* Splash. In Figure 1 we can see an illustrative example of these cymbals in a drum kit[2].

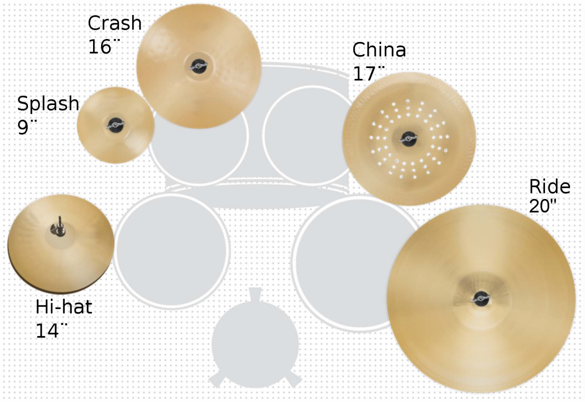


Fig. 1. Illustrative example with five different cymbal types in a drum kit

Hi-hat is a pair of cymbals mounted one above the other on a stand that is activated by the musician's foot. Splash are small cymbals, usually between 6" to 12" in diameter, with a very short decay. Often damped immediately after struck. Crash is a cymbal with a relatively short decay, used to accentuate musical phrases. China is a cymbal that has a slightly upturned edge and a conical bell that produces a sharp and *explosive* crash tone. Ride is a cymbal with higher dimensions primarily used in more popular music to execute rhythmic patterns.

We also divide each main category into subcategories according to the manner or region that the cymbals can be struck. For instance, China, Crash, Ride and Splash cymbals can be struck in the border region, called *Edge*. Ride cymbal also can be played in their raised center called *Bell*. The size of the bell determines the amount of overtones that will emanate from the instrument. Another possibility for Ride cymbals it is be struck in their *Body* (also called *Bow*), a region between the *Edge* and *Bell*. See Figure 2 to better understand.

For China and Crash cymbals we also consider the movement of *Roll*. Here Roll means a fast succession of single

[1] http://www.paiste.com/

[2] This figure was built with the support of tool Zildjian SoundLab available online at http://zildjian.com/soundlab

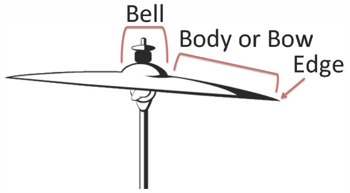


Fig. 2. Anatomy of traditional Turkish cymbal

notes on a cymbal at the edge region. In Splash cymbal we consider the movement of *Choke* where the cymbal is muffled with a hand after being struck. For Hi-hat we consider three subcategories: *Chick*, the sound produced by Hi-hats when closed with the foot; *Closed*, when the Hi-hat is struck in the edge and the bottom and top cymbal are together; *Open*, the sound produced after struck when the top and bottom cymbals are slightly separated.

Table I shows the distribution of each cymbal type and their subcategories in our evaluated dataset.

TABLE I. CLASS DISTRIBUTION OF CYMBAL DATASET

Cymbal Type	Characteristic	# Examples per Subclass	# Examples per Class (%)	Total
China	Edge Roll	51 48	99 (9.41)	1,052
Crash	Edge Roll	152 151	303 (28.80)	
Hi-hat	Chick Closed Open	75 77 75	227 (21.58)	
Ride	Bell Body Edge	107 113 124	344 (32.70)	
Splash	Choke Edge	36 43	79 (7.51)	

B. Features Extraction

We evaluate the most popular features extraction methods used in signal processing applications such as speech and musical instrument recognition. Briefly, these methods are responsible to change the original data representation in a more representative feature set for supervised learning algorithms. A description of each method is presented in the follow.

LSF. The Linear Spectral Frequencies [19] is a signal representation derived from Linear Predictive Coding (LPC) [20]. In LPC, a signal is represented as a linear combination of previously observed values according to Equation 1.

$$\hat{x}_k = \sum_{i=1}^{p} a_i x_{k-i} \tag{1}$$

where k is the time index and p is the number coefficients. The a_i coefficients are calculated in order to minimize the prediction error using covariance, auto-correlation or recursive least squares algorithms.

Equation 1 can be rewritten in the frequency domain with a z-transform [21]. In this way, a short segment of signal is assumed to be generated as the output of an all-pole filter $H(z) = \frac{1}{A(z)}$, where $A(z)$ is the inverse filter such that:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2}$$

As previously discussed, LSF is an alternative way to represent LPC coefficients. In order to calculate LSF, the inverse polynomial filter is decomposed into two polynomials $P(z)$ and $Q(z)$:

$$\begin{aligned} P(z) &= A(z) + z^{p+1}A(z^{-1}) \\ Q(z) &= A(z) - z^{p+1}A(z^{-1}) \end{aligned} \tag{3}$$

where $P(z)$ is a symmetric polynomial and $Q(z)$ is an anti-symmetric polynomial. The roots of $P(z)$ and $Q(z)$ determine the LSF coefficients. Thus, LSF can represent a large signal using a small number of coefficients. In this work we evaluate different amount of coefficients and we found empirically that 100 coefficients are adequate for our classification problem.

LFC and MFCC. The Linear-Frequency Cepstrum (LFC) and Mel-Frequency Cepstrum Coefficients (MFCC) [22] are a feature sets extracted from the cepstrum of the signal. The cepstrum is the inverse Fourier transform of log-magnitude spectrum of a signal. LFC uses the linear representation on the cepstrum and MFCC rescale the cepstrum based on known variation of the human ear's critical bandwidth with frequency called *mel* scale which replicates the human hearing perception. Equation 4 shows the conversion from frequency (f) to *mel*.

$$mel = 2595 \times log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

For many problems in signal processing such as speech recognition, MFCC are state-of-the-art for audio description. It is frequently assumed in the literature that in general 13 is a good choice for the amount of coefficients, for instance [3], [6], [23], [24]. However, as pointed in [25], a different amount of coefficients can achieve better results in some applications that deal with short duration signals. Thus, we evaluated a range of values between 10 and 50 and we chose 40 coefficients as feature descriptor for our problem. The same amount of coefficients is used for LFC.

Temporal. In Temporal set we consider 13 features extracted from temporal domain such as Short-time Energy, Magnitude Average, Root Mean Square, Mean, Temporal Centroid, Zero-crossing Rate, Interval, Complexity Estimate, Variance, Standard Deviation, Skewness, Kurtosis and Duration of signal.

Spectral. In Spectral set we consider some of basic features from Temporal set, but calculated in the spectral domain such as Energy, Spectral Centroid, Variance, Standard Deviation, Skewness, Kurtosis and Mean. We also extracted more 10 features as Fundamental Frequency, Inharmony, Tristimulus 1, Tristimulus 2, Tristimulus 3, Irregularity, Modified Irregularity, Spectral Flux, Roll-off and Flatness.

A more detailed discussion about how Temporal and Spectral features are calculated can be found in [26].

We make available in our website[3] both audio samples sorted into categories and subcategories as well as the extracted audio features.

V. Experimental Evaluation

In our experiments, we evaluate five traditional machine learning algorithms: Naive Bayes (NB) [27], C4.5 [28], Random Forest (RF) [29], k-Nearest Neighbor (KNN) [30] and Support Vector Machine (SVM) using Sequential Minimal Optimization [31].

For KNN we analyze different values for the parameter k (values between 1 to 15) and for SVM we evaluate different values for parameters c and γ (c varying from 1 to 3 and γ from 0.001 to 0.1) using the Grid Search technique [32]. We also varied the SVM's kernel in our experiments (Polynomial and Radial Basis Function – RBF).

Ten-fold cross-validation was used to partition the data into training and test sets. In other words, we break the data into 10 mutual sets of size approximately $n/10$, train the classifier on 9 sets and test on the remaining set. This process is repeated 10 times, each time with a different test set, and the mean test set accuracy is taken. We repeated this process ten times, randomizing the order of examples between two consecutive executions, i.e., we performed 10x10-fold cross-validation.

In the next sections we will better discuss our results for the classification of cymbal type and classification of cymbal type and the respective manner or region that the cymbal is struck.

We present our results using four main evaluation measures: Accuracy, Precision, Recall and F-Measure. Considering the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates from a confusion matrix, these measures are calculated according to Equations 5–8. A more detailed discussion about these measures can be found in [33].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{8}$$

A. Classification of Cymbal Type

Our first experiment consists in classify the signals according to the five possible cymbal types: China, Crash, Hi-hat, Ride and Splash. The accuracy results for different algorithms and feature sets can be seen in Table II. These values are the mean accuracy on a set of 10x10-fold cross-validation executions. In this table, we highlight in bold the best results achieved by an algorithm given a feature set and we underline the best result achieved by a feature set given an algorithm.

TABLE II. Accuracy results for cymbal type classification

Feature set	NB	C4.5	RF	KNN	SVM
LSF	80.90	88.05	93.08	93.59	**96.59**
MFCC	82.68	85.41	90.09	89.33	**94.01**
LFC	73.10	80.41	86.94	85.65	**91.41**
Temporal	70.28	86.51	**89.95**	82.27	85.85
Spectral	66.38	86.48	**89.38**	79.57	87.54

We note in Table II that the best result is achieved by the SVM classifier using the Line Spectral Frequencies, followed by the same classifier with MFCC feature set. In general, all classifiers have achieved their best results with these two feature sets. More specifically, C4.5, Random Forest, k-Nearest Neighbor and SVM have achieved their best results using LSF and the Naive Bayes algorithm with MFCC.

Each single execution of the 10-fold cross-validation procedure produced a confusion matrix where we can analyze the error between the classes. In Table III we show an example of one of 10 executions of the SVM classifier with the LSF feature set. In this table, we also show the Precision, Recall and F-Measure rates for each class.

TABLE III. Example of confusion matrix obtained by SVM classifier (kernel RBF, $c = 3.0$ and $\gamma = 0.01$) with LSF feature set for cymbal type classification

Actual	Predicted				
	China	Crash	Hi-hat	Ride	Splash
China	99	0	0	0	0
Crash	0	292	0	11	0
Hi-hat	0	0	225	2	0
Ride	0	21	1	322	0
Splash	0	1	0	0	78
Precisi•n	1.00	0.93	0.99	0.96	1.00
Recall	1.00	0.96	0.99	0.94	0.99
F-Measure	1.00	0.95	0.99	0.95	0.99

We can see in Table III that the classifier achieves very good results for China, Hi-hat and Splash cymbals. Basically, the mistakes are concentrated in Ride and Crash cymbals. In practical terms, these errors are not a great problem. Due to the sound similarity between these two cymbals, mainly between Crash with larger dimensions and Ride in general, many drummers use only Crash cymbals in their drum sets.

B. Classification of Characteristic or Movement

In our second experiment, the main goal is to achieve a more informative response. In addition to classify the cymbal type, we identify how the sound of this cymbal was produced. This specificity is especially useful for musical drum transcription applications. The possibilities considered for each cymbal were previous presented in Section IV-A. The general results in terms of accuracy are presented in Table IV.

TABLE IV. Accuracy results for classification of cymbal type and their characteristic

Feature set	NB	C4.5	RF	KNN	SVM
LSF	78.45	71.53	80.09	74.63	**86.49**
MFCC	65.23	63.90	69.72	56.40	**74.93**
LFC	56.96	57.55	66.53	58.67	**72.91**
Temporal	68.07	82.17	**86.02**	72.93	81.54
Spectral	78.59	80.77	**85.16**	73.18	84.36

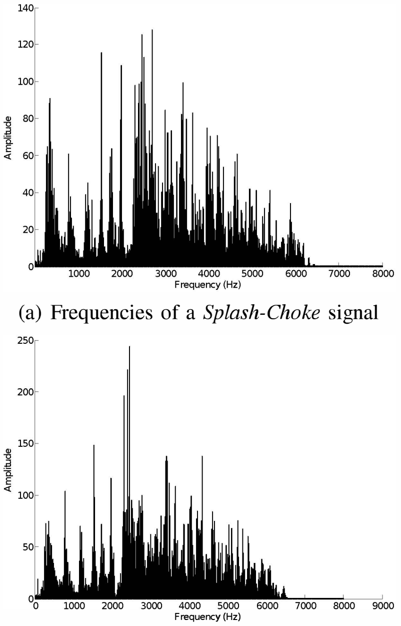
[3] http://sites.labic.icmc.usp.br/vsouza/paiste_dataset/

TABLE V. EXAMPLE OF CONFUSION MATRIX OBTAINED BY SVM CLASSIFIER (KERNEL RBF, $c = 3.0$ AND $\gamma = 0.01$) WITH LSF FEATURE SET FOR CLASSIFICATION OF CYMBAL TYPE AND THEIR CHARACTERISTIC

Actual	**Predicted**											
	CE	CR	CrE	CrR	HCh	HCl	HOp	RBe	RBo	REd	SC	SE
China-Edge (CE)	43	8	0	0	0	0	0	0	0	0	0	0
China-Roll (CR)	14	34	0	0	0	0	0	0	0	0	0	0
Crash-Edge (CrE)	1	0	130	10	0	0	0	0	0	11	0	0
Crash-Roll (CrR)	0	0	22	129	0	0	0	0	0	0	0	0
Hi-hat-Chick (HCh)	0	0	0	0	72	0	0	0	0	0	0	0
Hi-hat-Closed (HCl)	0	0	0	0	1	75	1	0	0	0	0	0
Hi-hat-Open (HOp)	0	0	0	0	0	0	74	0	1	0	0	0
Ride-Bell (RBe)	0	0	0	0	0	0	1	104	0	2	0	0
Ride-Body (RBo)	0	0	0	0	0	0	0	0	113	0	0	0
Ride-Edge (REd)	0	0	16	1	0	0	0	3	1	103	0	0
Splash-Choke (SC)	0	0	2	0	0	0	0	0	0	0	7	27
Splash-Edge (SE)	0	0	2	0	0	0	0	0	0	0	18	23
Precisi●n	0.74	0.81	0.76	0.92	0.99	0.96	0.97	0.97	0.98	0.89	0.28	0.46
Recall	0.84	0.71	0.85	0.85	0.96	0.97	0.99	0.97	1.00	0.83	0.19	0.53
F-Measure	0.79	0.76	0.80	0.89	0.97	0.97	0.98	0.97	0.99	0.86	0.23	0.49

Again, we note that the SVM classifier with LSF achieved the best result with 86.49% of accuracy. However, unlike the first experiment, Temporal features also show good results.

In Table V we show the confusion matrix of one of 10 executions of SVM classifier with the LSF feature set for the classification of cymbal type and their respective characteristic. We observe that this new classifier inherits the errors from the first classifier presented, with mistakes concentrated in Crash and Ride cymbals. This type of error is expected. However, the higher error rate occurs with Splash cymbal when the *Choke* movement is performed. Thus, the Splash-Choke and Splash-Edge classes have achieved very low results with 0.23 and 0.49 of F-Measure. These low F-Measure results were achieved with LSF. This feature set basically extracts characteristics from the frequency domain. Thus, it is expected a difficulty in distinguish sounds from the same cymbal type. For instance, consider the example of signals of *Splash-Choke* and *Splash-Edge* in frequency domain showed in Figure 3.

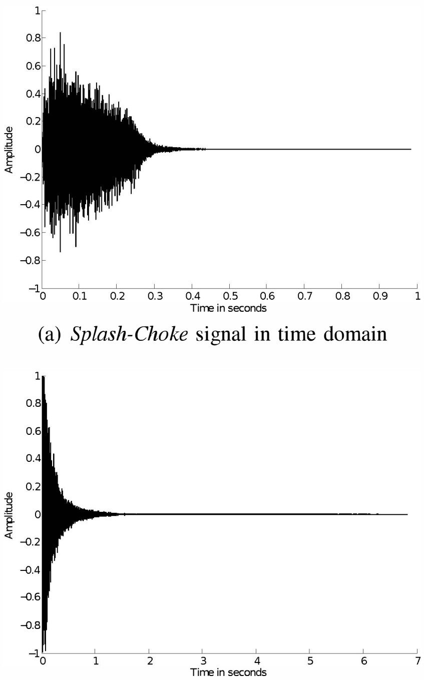


(a) Frequencies of a *Splash-Choke* signal

(b) Frequencies of a *Splash-Edge* signal

Fig. 3. Two different examples of Splash signals in frequency domain

We can see in Figure 3 that both signals are very similar and justifies the classifier's errors. On the other hand, if we analyze the same signals in time domain, as presented in Figure 4, we can note clearly that sounds generated in the regions of *Choke* and *Edge* of Splash cymbals has significant differences in the ADSR envelope features: attack, decay, sustain, and release.



(a) *Splash-Choke* signal in time domain

(b) *Splash-Edge* signal in time domain

Fig. 4. The same Splash signals of Figure 3 in time domain

The ADSR envelope have a great effect on the instrument's sonic character. Basically, attack is the time it takes for the note to reach the maximum level. Decay is the time it takes for the note to go from the maximum level to the sustain level. Sustain is the level while the note is held. Release is the time it takes for the note to fall from the sustain level to zero (silence) when released. An illustrative example that represents the ADSR envelope is shown in Figure 5.

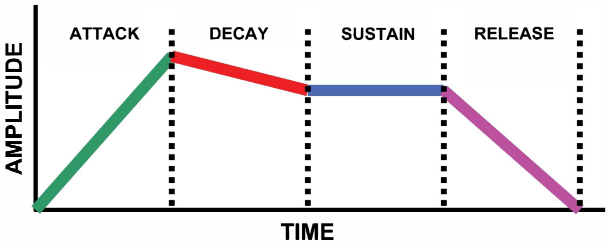


Fig. 5. Representation of Attack, Decay, Sustain, Release – ADSR envelope

Such an observation about the signals in time domain and the good results achieved by the Temporal feature set as shown in Table IV, are indicatives that LSF together with Temporal can help to improve the results. Thus, we merge both feature sets and evaluate the results for different algorithms. Table VI shows the achieved results and confirms our intuition. The *improvement* row refers to the accuracy improvement in percentage compared to the best result showed in Table IV achieved by SVM with LSF (86.49%). Indeed, the SVM classifier with LSF and Temporal feature sets improves the results in more than 5% of accuracy, i.e., more 53 examples correctly classified in a total of 91.54% of accuracy.

TABLE VI. ACCURACY RESULTS FOR CLASSIFICATION OF CYMBAL TYPE AND THEIR CHARACTERISTIC CONSIDERING LSF AND TEMPORAL FEATURE SETS TOGETHER

Feature set	**NB**	**C4.5**	**RF**	**KNN**	**SVM**
LSF + Temporal	85.68	87.36	89.36	83.77	**91.54**
Improvement	7.09	5.19	3.34	9.14	5.05

An example of the confusion matrix considering the SVM classifier and the LSF together with Temporal feature sets is shown in Table VII. We can note that with LSF and Temporal feature sets together, the classes *Splash-Choke* and *Splash-Edge* have improved considerably their results. In addition, other classes such as *China-Edge, China-Roll, Crash-Edge, Crash-Roll, Hi-hat-Open, Ride-Bell*, and *Ride Edge* also have improved slightly their results. These gains can be observed in the three measures considered (Precision, Recall and F-Measure).

VI. Conclusions and Future Work

In this paper we present the task of drum's cymbal classification. We classify the cymbal in categories of type and the manner that they are struck. To the best of our knowledge, no other work in literature has investigated the classification of indefinite pitch percussion sounds with this specificity. We believe that our results can be useful for audio professionals and for a more detailed automatic drum transcription as well as for other related applications.

We show experimentally that the SVM algorithm with LSF and Temporal features together has presented a good discriminative power (above 90% of accuracy on both discriminative levels discussed in the paper). As MFCC is many times considered the state-of-the-art method for audio description, we encourage the comparative evaluation against the LSF method in similar scenarios to ours.

In future research we plan to explore the association of subjective labels to cymbals by multi-label algorithms. In music, it is common the use of more subjective characteristics. For instance, drummers can say that a specific cymbal sound it is at the same time *dark, warm, broad, smoky, lively, bright* and others. In addition to explore different multi-label classifiers and good feature descriptors, we need relabel our data using a new set of labels. These labels can be retrieved from textual data in web forums about drums or achieved by means of questionnaires carried out for drummers about the sound samples.

Another direction is explore different signal representations such as spectrogram, chromagram or recurrence plots for cymbal sounds. From these representations, it is possible to extract interesting and less explored features such as based in texture as shown in [34], [35].

Acknowledgment

The authors would like to thank Fundação de Amparo a Pesquisa e Desenvolvimento do Estado de São Paulo (FAPESP) for financial support in the award number 2011/17698-5.

References

[1] P. Herrera Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3–21, 2003.

[2] G. Tzanetakis, "Anssi klapuri, manuel davy, eds: Signal processing methods for music transcription," *Computer Music Journal*, vol. 32, no. 4, pp. 86–88, 2008.

[3] P. Herrera, A. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques," in *International Conference on Music and Artificial Intelligence (ICMAI)*, 2002, pp. 69–80.

[4] J. Paulus, "Acoustic modelling of drum sounds with hidden markov models for music transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2006, pp. 241–244.

[5] K. Yoshii, M. Goto, and H. G. Okuno, "Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 333–345, 2007.

[6] S. Scholler and H. Purwins, "Sparse approximations for drum sound classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 933–940, 2011.

[7] S. Cavaco and H. Almeida, "Automatic cymbal classification using non-negative matrix factorization," in *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2012, pp. 468–471.

[8] J. Paulus and A. Klapuri, "Model-based event labeling in the transcription of percussive audio signals," in *Proceedings of Digital Audio Effects Workshop (DAFx)*, 2003.

[9] D. Van Steelant, K. Tanghe, S. Degroeve, B. De Baets, M. Leman, J.-P. Martens, and T. De Mulder, "Classification of percussive sounds using support vector machines," in *Proceedings of the annual machine learning conference of Belgium and The Netherlands*, 2004.

[10] O. Gillet and G. Richard, "Automatic transcription of drum loops," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2004, pp. 266–269.

[11] E. Zwicker and U. T. Zwicker, "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system," *Journal of the Audio Engineering Society*, vol. 39, no. 3, pp. 115–126, 1991.

[12] J. J. Harrison and A. J. Hill, "A scientific approach to microphone placement for cymbals in live sound," *Institute of Acoustics*, vol. 35, pp. 219–228, 2013.

TABLE VII. Example of confusion matrix obtained by SVM classifier (kernel RBF, $c = 3.0$ and $\gamma = 0.01$) with LSF and Temporal feature sets for classification of cymbal type and their characteristic

Actual	**Predicted**											
	CE	CR	CrE	CrR	HCh	HCl	HOp	RBe	RBo	REd	SC	SE
China-Edge (CE)	44	7	0	0	0	0	0	0	0	0	0	0
China-Roll (CR)	12	36	0	0	0	0	0	0	0	0	0	0
Crash-Edge (CrE)	1	0	132	9	0	0	0	0	0	10	0	0
Crash-Roll (CrR)	0	0	17	132	0	0	0	0	0	1	1	0
Hi-hat-Chick (HCh)	0	0	0	0	72	3	0	0	0	0	0	0
Hi-hat-Closed (HCl)	0	0	0	0	2	75	0	0	0	0	0	0
Hi-hat-Open (HOp)	0	0	0	0	0	0	74	0	0	0	1	0
Ride-Bell (RBe)	0	0	0	0	0	0	0	107	0	0	0	0
Ride-Body (RBo)	0	0	0	0	0	0	0	0	113	0	0	0
Ride-Edge (REd)	0	0	15	1	0	0	0	1	1	106	0	0
Splash-Choke (SC)	0	0	0	0	0	0	0	0	0	0	34	2
Splash-Edge (SE)	1	0	4	0	0	0	0	0	0	0	0	38
Precisi●n	0.76	0.84	0.79	0.93	0.97	0.96	1.00	0.99	0.99	0.91	0.94	0.95
Recall	0.86	0.75	0.87	0.87	0.96	0.97	0.99	1.00	1.00	0.85	0.94	0.88
F-Measure	0.81	0.79	0.82	0.90	0.97	0.97	0.99	0.99	0.99	0.88	0.94	0.92

[13] A. Chaigne, C. Touze, and O. Thomas, "Nonlinear vibrations and chaos in gongs and cymbals," *Acoustical science and technology*, vol. 26, no. 5, pp. 403–409, 2005.

[14] T. D. Rossing, "Acoustics of percussion instruments: Recent progress." *Acoustical Science and Technology*, vol. 22, no. 3, pp. 177–188, 2001.

[15] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," in *Audio Engineering Society Convention 115*. Audio Engineering Society, 2003, pp. 1–14.

[16] P. Herrera Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal processing methods for music transcription*, 2006, pp. 163–200.

[17] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 68–80, 2006.

[18] ——, "Musical instrument recognition by pairwise classification strategies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1401–1412, 2006.

[19] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[20] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[21] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*. Prentice-hall, vol. 2.

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[23] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Automatic Speech Recognition: Challenges for the new Millenium*, 2000, pp. 1–8.

[24] D. F. Silva, V. M. A. Souza, G. E. A. P. A. Batista, and R. Giusti, "Spoken digit recognition in portuguese using line spectral frequencies," in *Proceedings of the 13th Ibero-American Conference on Artificial Inteligence (IBERAMIA)*, 2012, pp. 241–250.

[25] D. F. Silva, V. M. A. Souza, and G. E. A. P. A. Batista, "A comparative study between MFCC and LSF coefficients in automatic recognition of isolated digits pronounced in portuguese and english," *Acta Scientiarum. Technology*, vol. 35, no. 4, pp. 621–628, 2013.

[26] T. H. Park, P. Lansky, and P. Cook, *Towards automatic musical instrument timbre recognition*. Princeton University, 2004.

[27] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995, pp. 338–345.

[28] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[31] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.

[32] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003.

[33] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[34] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, F. Gouyon, and J. G. Martins, "Music genre classification using lbp textural features," *Signal Processing*, vol. 92, no. 11, pp. 2723–2737, 2012.

[35] V. M. A. Souza, D. F. Silva, and G. E. A. P. A. Batista, "Extracting texture features for time series classification," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2014, pp. 1425–1430.