TOWARD UNDERSTANDING EXPRESSIVE PERCUSSION THROUGH CONTENT BASED ANALYSIS

Matthew Prockup, Erik M. Schmidt, Jeffrey Scott, and Youngmoo E. Kim
Music and Entertainment Technology Laboratory (MET-lab)
Electrical and Computer Engineering, Drexel University
{mprockup,eschmidt,jjscott,ykim}@drexel.edu

ABSTRACT

Musical expression is the creative nuance through which a musician conveys emotion and connects with a listener. In un-pitched percussion instruments, these nuances are a very important component of performance. In this work, we present a system that seeks to classify different expressive articulation techniques independent of percussion instrument. One use of this system is to enhance the organization of large percussion sample libraries, which can be cumbersome and daunting to navigate. This work is also a necessary first step towards understanding musical expression as it relates to percussion performance. The ability to classify expressive techniques can lead to the development of models that learn the the functionality of articulations in patterns, as well as how certain performers use them to communicate their ideas and define their musical style. Additionally, in working towards understanding expressive percussion, we introduce a publicly available dataset of articulations recorded from a standard four piece drum kit that captures the instrument's expressive range.

1. INTRODUCTION

In music, it is the human component of expression that imparts emotion and feeling within a listener. Expression relates to the nuances in technique that a human performer imparts on a piece of music. Musicians creatively vary timing, dynamics, and timbre of the musical performance, independent from the score, in order to communicate something of deeper meaning to the listener [1]. For example, a musician can alter tempo or change dynamics slightly to impart tension or comfort. Similarly, they can alter the timbre of their instrument to create different tonal colors. All of these parameters add an additional level of intrigue to the written pitches, rhythms, and dynamics being performed.

In studying percussion, one of the fundamental ways of communicating a musical idea is through expressive *articulation*. Differences in articulation are created by the creative combination of dynamics and excitation timbre. This simple relationship is outlined in Figure 1. There are an almost infinite number of ways that a percussionist can strike a drum. While the strike itself is restricted to being a single discrete event, there exists a vast range of articulations that make each of those seemingly discrete actions sit in a continuous and highly dimensional space.

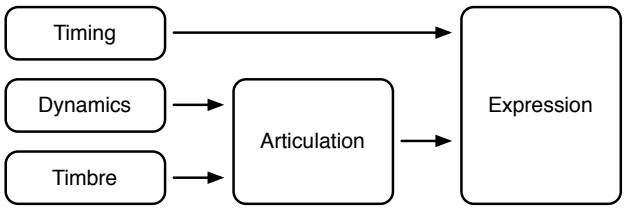

Figure 1. Expression: Creative alterations in timing, dynamics, and instrument timbre can define a musician's expressive style.

In percussion, there are four main techniques of excitation: strikes, rim shots, cross sticks, and buzz strokes. An explanation of these techniques is outlined in Table 1. This simple set of excitation techniques become the building blocks of the standard *rudiments* that define most aspects of percussion music [2]. Each expressive articulation has meaning in the context of a rudiment, and many individual performers have unique ways of expressing and combining them. This defines their style and identity as a musician. In this initial work, we seek to quantify and understand differences in excitation techniques. It is important in the context of percussion that a bottom up approach be taken to expressive performance analysis. Percussion performance is built on the rudimentary combination of unique articulations, so this is a logical place to start. In the music information retrieval community, it has been a large aspect of percussion performance and expression that has been ignored.

In working towards this understanding of expressive percussion, we have compiled a comprehensive new public dataset of expressive samples recorded from a standard four piece drum kit. The dataset includes samples varied by intensity of stroke (staccato vs legato), height of stroke, and strike position over a variety of excitation techniques for each instrument of the drum kit. Using this dataset we train a simple four class support vector machine (SVM) to distinguish these expressive articulations both depen-

Articulation	Description
Strike	The drumhead is struck with the tip of the stick.
Rim Shot	Both the drumhead and rim are struck with the tip and shaft of the stick simultaneously.
Cross Stick	The butt of the stick strikes the rim while the tip rests on the head.
Buzz Stroke	The stick is pressed into the drum to create multiple, rapid strokes.

Table 1. Excitation Techniques: There are four basic drum excitation techniques.

dent on and independent of percussion instrument type. In the context of this paper, we will investigate the three drums commonly struck with sticks (snare drum, rack tom, and floor tom) and the four excitations that become the building blocks of rudiments. Excitation classification is only a small aspect of percussion expression, but the ability to recognize these differences in articulation is a necessary first step in understanding percussion performance as a whole.

2. BACKGROUND

There are a few areas of research tangentially related to expressive percussion performance. The first and most widely studied is the task of instrument identification. Earlier studies in instrument recognition have focused mainly on the ability to classify a wide range of traditional instrument tones, but more recently, a greater effort has been made to classify instruments specific to the realm of percussion. In [3], a set of systems using a wide range of feature selection and classification techniques performed well at discriminating percussion instruments. However, this study only took into account a standard drum strike and purposely did not include alternative articulations, such as rim shots or buzz strokes.

Some studies take the instrument identification approach a step further and attempt to transcribe drum patterns. One such transcription study presented in [4] used non-negative spectrogram factorization and onset detection techniques in order to separate drum sounds and classify them as either a snare drum, bass drum, or hi-hat. This shows promise in the ability to retrieve drum sounds directly from patterns. In [5], Battenberg and Wessel used deep learning approaches in order to learn beat sequence timings of the snare drum, bass drum, and hi-hat in different drum patterns. Understanding a drum's context within a performance can lead to models that can inform musical style. This was a step in the right direction for the analysis of percussion expression.

There has also been an evolving volume of work studying musical performance analysis and expression specifically. Mion and Poli in [1] stated that musical expression is best represented with score independent descriptors that model intricacies in timing, dynamics, and timbre. They showed that a simple set of features can be used to capture and classify the expressive intent of a performer in both affective and sensorial domains. Other work in music expression focuses on the intricacies of specific instruments. In [6], an analysis-by-synthesis experiment was performed to model, synthesize, and evaluate the expressive characteristics of a clarinet performance. The authors identified feature dynamics that relate to expressive performance. They then forced the dynamic features to be static, creating a less expressive re-synthesis. A listening test was then performed which asked if subjects preferred the original or altered recordings. Results from the test showed that listeners preferred the original musically expressive performance. It also showed that expression is captured in the evolution of features over time, and removing this aspect effectively removes musical expression. This demonstrated that the dynamic nature of instrument timbre is an important aspect of music expression. In order to capture feature dynamics, simple polynomial expressions can be fit to the time varying process. This provides a compact representation of sequential data in both the time and frequency domains [7].

A vast majority of prior work in musical expression analysis has revolved around understanding the timbral characteristics of pitched instruments. A detailed analysis of expressive percussion is also necessary, yet it is largely ignored. However, some sparse examples of these studies do exist. The work in [8] focuses on snare drum expression and attempts to distinguish playing position on the head as well as excitation techniques, such as using brushes or playing a rim shot. These experiments, however, were very limited in scope, with models being only applicable to one drum. Additionally, all training and testing examples were performed at a single volume and intensity level.

In this paper, we perform the task of percussion articulation classification similar to the work found in [8]. In our study however, it is important for the models to generalize over multiple pieces of the drum kit. Secondly, our models incorporate additional excitation techniques (buzz strokes and cross stick strokes) as well as a dataset containing many different ways of performing these articulations. Using compact representations of timbral characteristics over time, we train classifiers to distinguish excitation techniques independent of drum, stick height, intensity of stroke, and head strike position.

3. DATASET OF EXPRESSIVE PERCUSSION

In domains outside of percussion, there exist large datasets that can be used for expressive performance analysis. A comprehensive, well-labeled set of expressive percussion samples is less common. The presented work makes use of a newly recorded dataset that encompasses a vast array of percussion performance expressions on a standard four piece drum kit. In the context of this paper, only the snare drum, rack tom, and floor tom samples are used. Each drum used has samples that span the following range:

- stick heights: 8cm, 16cm, 24cm, and 32cm
- stroke intensities: light, medium, heavy

Feature Names	Feature Abbreviation	Feature Description	Source
RMS energy	RMS	root-mean-squared energy	n/a
roughness	R	energy of beating frequencies	[9]
brightness	B	description of spectral brightness	[9]
2 bin ratio (bottom half)	SRA	ratio of spectral energy below 1000Hz to the full spectrum	[1]
3 bin ratio (low)	SRL	ratio of spectral energy below 534Hz to the full spectrum	[1]
3 bin ratio (med)	SRM	ratio of spectral energy between 534Hz and 1805Hz to the full spectrum	[1]
3 bin ratio (high)	SRH	ratio of spectral energy above 1805Hz to the full spectrum	[1]

Table 2. Basic Features: Single dimensional time and frequency domain features are used as the basis for the evolution features.

- strike positions: center, halfway, edge
- articulations: strike, rim shot, buzz stroke, cross stick

This subset includes 1804 individual examples across the four articulations over the three drums. Additionally, there are at least 4 examples of each expressive combination. Recordings include samples with the snare wires both touching (snares on) and not touching (snares off) the bottom head of the snare drum. The division of sample variety is not completely uniform across the entire set, but it was designed to allow for the most complete coverage of each instrument's expressive range. That being said, no one combination of expressive parameters vastly outweighs another and all are adequately represented.

The full dataset also includes a complete array of expressive bass drum, hi-hat, and cymbal samples as well. Each articulation example has monophonic and stereo versions with multiple mixes using direct (attached) and indirect (room) microphone positioning techniques. This is the first publication where this dataset appears and it can be made freely available to others upon request.

4. PREDICTING EXPRESSIVE ARTICULATION

In expressive performance, the evolution of timbre over time is an important component on both a micro and macro level. This work investigates expression at the micro level by attempting to model the evolution of percussion articulations. Using the sequential evolution of features derived from time domain and frequency domain components of the signal, a set of classifiers is trained to predict percussion articulations within subsets containing only individual drums (only snare, only rack tom, etc.) as well as within the superset of all drum samples.

4.1 Feature Design

The aural differences in percussion articulations are defined by the short time evolution of their spectral components. For example, a buzz stroke evolves very differently than a rim shot. These differences are apparent in both their time domain and frequency domain characteristics. In order to capture this evolution, a set of compact features was implemented that model the envelope of single dimensional features over time. This compact representation is derived from the coefficients of a polynomial fit to the time varying feature data similar to [7]. This compact polynomial representation was calculated for the features outlined in Table 2. Descriptions of the new polynomial coefficient features are described in Table 3.

Feature Names	Feature Description
RMS_3 RMS_6	3^{rd} and 6^{th} order coefficients of RMS
R_3 R_6	3^{rd} and 6^{th} coefficients of R
B_3 B_6	3^{rd} and 6^{th} coefficients of B
SR_3 SR_6	3^{rd} and 6^{th} aggregated coefficients of SRA, SRL, SRM, and SRH

Table 3. Evolution Features: New features are derived from the coefficients of polynomials fit to the the single dimensional features in Table 2 over time.

Figure 2 shows the time evolution of selected features and their polynomial representations for a snare drum across each of the articulation examples. It is easy to qualitatively discriminate the differences in shape for each of the articulations. Polynomials fit to the feature data are able to capture this shape in a compact manner. It was found in early experimentation that the third and sixth degree polynomial fits were optimal for representation. In order to evaluate the salience of these newly implemented features, Mel-Frequency Cepstral Coefficients (MFCCs) and their first and second derivatives were also used in the classification tasks for comparison.

4.2 Experiments

The main focus of the work presented is to classify the excitation techniques of expressive drum strike articulations. The articulations observed and their descriptions are shown in Table 1. Using the polynomial coefficient features from Table 3, a four class support vector machine (SVM) using a radial basis function (RBF) kernel was trained to discriminate excitation. In all experiments, five-fold cross validation was performed for both parameter tuning and training/testing. The classification task was run for each drum individually as well as for all drums in combination. This tested the effectiveness of the system to understand expression on individual drums as well as throughout the entire drum kit. For example, in a robust system a rim shot should be classified as such regardless of the instrument on which it was performed. In order to compare the effectiveness of each of the new features, the classification task was also performed using the means of the MFCCs and their first and second derivatives over the duration of the sample.

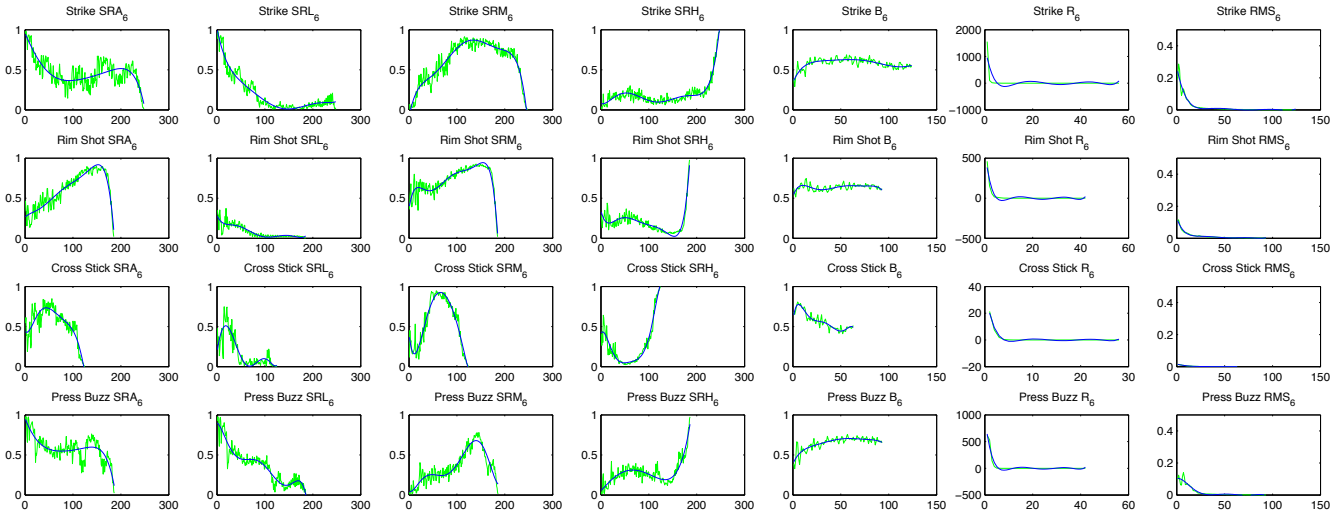

Figure 2. Feature Evolution Example: Sixth order polynomials are fit to the temporal feature data of four snare drum articulations.

The first experiment involved classifying excitation on the each drum individually. Features were used both alone and in aggregation. In order to aggregate the features, each dimension was normalized to have zero mean and unit variance. The testing data was transformed using the mean and variance derived from the training data. This allowed each feature to be simply concatenated for training and testing. The raw features and projections via a principal components analysis (PCA) were also explored, but in practice, the simple normalization transformation yielded the best results. The second experiment classified excitation over the set of all drum samples. Again, the features were used both individually and in aggregation with the simple normalization. In both experiments, the new features and their combinations were also used in conjunction with MFCCs. This MFCC aggregation shows their ability to add time domain information to an already salient, yet static, feature and improve its performance.

4.3 Results

The first experiment classifies excitation for each drum independently using the features individually as well as in selected aggregations. Table 4 shows the accuracies for the features individually. MFCCs averaged over the example are the best single performing feature for both the snare drum and rack tom. The floor tom, however, shows better performance with the 3rd and 6th order polynomial coefficients of the spectral ratios (SR_3 SR_6) than it does with the MFCCs. While standard MFCCs do not take into account any information about time evolution, each articulation does have an inherently different average timbre. Because MFCCs are designed to provide an estimate of the spectral envelope and capture this timbre, they perform reasonably well. However, when the samples have a greater length and therefore a longer timbre evolution, such as that of a floor tom, MFCCs start to degrade in performance while some of the evolution features start to improve.

Individual Feature	Snare	Rack Tom	Floor Tom
MFCC	$\mathbf{0.956 \pm 0.012}$	$\mathbf{0.914 \pm 0.037}$	0.872 ± 0.027
Δ MFCC	0.771 ± 0.015	0.661 ± 0.029	0.834 ± 0.015
Δ^2 MFCC	0.646 ± 0.031	0.544 ± 0.061	0.637 ± 0.020
SR_3	0.897 ± 0.016	0.835 ± 0.017	$\mathbf{0.907 \pm 0.045}$
SR_6	0.776 ± 0.023	0.838 ± 0.033	0.896 ± 0.025
B_3	0.736 ± 0.032	0.846 ± 0.031	0.859 ± 0.036
B_6	0.713 ± 0.013	0.755 ± 0.032	0.845 ± 0.009
R_3	0.407 ± 0.017	0.670 ± 0.017	0.523 ± 0.022
R_6	0.514 ± 0.021	0.822 ± 0.017	0.578 ± 0.050
RMS_3	0.637 ± 0.013	0.696 ± 0.039	0.795 ± 0.039
RMS_6	0.773 ± 0.014	0.893 ± 0.030	0.845 ± 0.018

Table 4. Classification Accuracies: Excitation techniques were classified using each feature on each drum individually.

Table 5 shows the performance of features in combination on the individual drums. The feature combinations with the highest classification accuracies for each drum are displayed along with the best performing individual features for comparison. In all cases, the aggregated feature combinations had a higher classification accuracy than each of the best performing individual features. This shows that combining an estimation of general timbre with certain features that capture that timbre's evolution can improve classification accuracy. In Table 5 only the top five performing feature combination accuracies for each drum are shown. Those that appear in multiple lists show they are better at generalizing over the different drum types. The 6^{th} order brightness feature in combination with MFCCs (B_6 MFCC) was the only aggregation to appear within the top five best performing combinations over all three drum types.

In the second experiment, a single classifier was trained on articulation samples from all three drums. The classifiers were again trained on each feature individually and in combination. The accuracies for the classification of percussion articulations, independent of drum, are shown in Table 6. In the classification of excitation over the superset

Feature Aggregation	Snare Drum	Rack Tom	Floor Tom
SR_3 R_3 B_3 MFCC	**0.987 ± 0.001**	-	0.982 ± 0.010
SR_3 B_3 MFCC	0.982 ± 0.005	-	0.972 ± 0.007
B_3 MFCC	0.982 ± 0.007	0.956 ± 0.013	-
B_6 MFCC	0.978 ± 0.005	0.963 ± 0.015	0.974 ± 0.019
SR_3 R_3 MFCC	0.977 ± 0.012	-	-
SR_6 R_6 B_6 MFCC	-	**0.9712 ± 0.009**	0.982 ± 0.014
SR_6MFCC	-	0.955 ± 0.020	-
R_6 MFCC	-	0.955 ± 0.012	-
SR_6 B_6 MFCC	-	-	**0.984 ±0.005**
Best Individual	0.956 ± 0.012 (MFCC)	0.914 ± 0.037 (MFCC)	0.907 ± 0.045 (SR_3)

Table 5. Classification Accuracies: Excitation techniques were classified using selected feature aggregations on each drum individually. Results are shown for the top five performing features on each drum. Feature combinations that are outside the top five best performing aggregations for a single drum type are marked with '-'.

of all drums, MFCCs were shown to be the best performing feature. However, when the polynomial envelope features were used in combination with MFCCs, accuracy was again improved. The 6^{th} order brightness feature in combination with MFCCs (B_6 MFCC) was the best performing feature for over the superset of all drums. This is likely due to the fact that this combination was also the only one contained within the top performing combinations of all individual experiments from Table 5.

Feature	All Drums
MFCC	**0.930 ± 0.011**
Δ MFCC	0.745 ± 0.021
Δ^2 MFCC	0.534 ± 0.016
SR_3	0.847 ± 0.010
SR_6	0.744 ± 0.020
B_3	0.734 ± 0.024
B_6	0.719 ± 0.018
R_3	0.498 ± 0.020
R_6	0.514 ± 0.006
RMS_3	0.731 ± 0.017
RMS_6	0.590 ± 0.008
B_6 MFCC	**0.972 ± 0.004**
SR_3 R_3 B_3 MFCC	0.969 ± 0.011
SR_6 B_6 MFCC	0.967 ± 0.008
SR_6 R_6 B_6 MFCC	0.965 ± 0.006
SR_3 B_3 MFCC	0.963 ± 0.004

Table 6. Classification Accuracies: Excitation techniques were classified using features individually and in aggregation over the superset of all drum types.

In all cases, for each individual drum and the superset of all drums, MFCCs performed rather well on their own. However, they do not take into account any information regarding the temporal evolution of the signal. The derivatives of MFCCs were also used, but they provide only a static picture of the amount of change present when averaged over the example. They still lack information as to how those changes evolve. Additionally, in the presented experiments, MFCCs were shown to be better at modeling articulations than were their derivatives. However, by using the polynomial coefficients of simple time varying features along with standard MFCCs, the system was able to gain temporal context, leading to better performance.

5. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, it was shown that the coefficients of polynomials fit to model feature evolution can provide a compact representation with the ability to quantify percussive articulation. These features in conjunction with popular features, such as MFCCs, can improve performance by adding temporal context. In this paper, we also introduced a new comprehensive dataset of expressive percussion articulations. This presented work only scratches the surface of this dataset's applicability to problems involving expression in musical performance. Classifying articulation is a small, yet very necessary step in the understanding of percussion performance and expression in general. Moving forward, more work must be done towards understanding the micro and macro evolution of expression.

On the micro level, this work can be expanded upon by using more sophisticated systems to improve the modeling of feature evolution. It was shown in [10] that linear dynamical systems (LDS) are a compact way of representing and synthesizing pitched percussive instrument tones. This introduces the possibility of training an LDS for each articulation example and training a classifier that uses system parameters as features. Secondly, an LDS is a generative model, so it may also be possible generate or alter learned sets of percussive articulation. Understanding this micro evolution can greatly assist in the navigation and organization of large humanly expressive sample libraries, which are usually cumbersome for percussion instruments.

In future work, we look to model not only the micro evolution, but the macro evolution of expression as well. If we are able to classify percussion articulations, we can look further into its meaning by developing models that learn the functionality of articulation in patterns and performance. The articulation classification along with statistics of their usage, dynamics, and time onsets can lead to models that contain information about human playing style. This performance style can be used to model individual percussionists or larger populations of similar percussionists. With these performance models in conjunction with the ability to classify articulation, we can investigate the possibility of expressive performance generation using unlabeled sets of any custom sample library that a producer or composer wishes to use. This may seem like a lofty goal in relation to this work's present state, and in most respects, it is. However, expressive articulation is one of the most important parameters of a percussionist's performance. The ability to classify expressive excitation, independent of percussion instrument, is the necessary first step towards understanding the unique intricacies and nuances of percussion performance and its relation to human expression in general.

6. ACKNOWLEDGMENTS

The authors would like to thank the Music Industry Department of Drexel University's College of Media Arts and Design for their support and assistance in the recording, mixing and organization of the expressive percussion sam-

ple library. It was with their help that we were able to create a comprehensive, high quality, labeled audio dataset of expressive percussion.

7. REFERENCES

[1] L. Mion and G. D. Poli, "Score-independent audio features for description of music expression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 458–466, 2008.

[2] M. Goldenberg, *Modern school for snare drum*. Hal Leonard, 1955.

[3] P. Herrera, A. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques," in *Music and Artificial Intelligence*, vol. 2445 of *Lecture Notes in Computer Science*, pp. 69–80, Springer Berlin Heidelberg, 2002.

[4] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proceedings of the 13th European Signal Processing Conference*, p. 4, 2005.

[5] E. Battenberg and D. Wessel, "Analyzing drum patterns using conditional deep belief networks," in *Proceedings of the International Conference on Music Information Retrieval*, 2012.

[6] M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad, "Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance," *Music Perception*, vol. 28, no. 3, pp. 265–278, 2011.

[7] M. Lagrange, M. Raspaud, R. Badeau, and G. Richard, "Explicit modeling of temporal dynamics within musical signals for acoustical unit similarity," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1498–1506, 2010.

[8] A. Tindale, A. Kapur, G. Tzanetakis, and I. Fujinaga, "Retrieval of percussion gestures using timbre classification techniques," in *Proceedings of the International Conference on Music Information Retrieval*, pp. 541–544, 2004.

[9] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *International Conference on Digital Audio Effects*, pp. 237–244, 2007.

[10] E. M. Schmidt, R. V. Migneco, J. J. Scott, and Y. E. Kim, "Modeling musical instrument tones as dynamic textures," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 329–332, IEEE, 2011.