Drum Stroke Computing: Multimodal Signal Processing for Drum Stroke Identification and Performance Metrics

Jordan Hochenbaum[1, 2]
New Zealand School of Music[1]
PO Box 2332
Wellington 6140, New Zealand
jhochenbaum@calarts.edu

Ajay Kapur[1,2]
California Institute of the Arts[2]
24700 McBean Parkway
Valencia CA, 91355
akapur@calarts.edu

ABSTRACT

In this paper we present a multimodal system for analyzing drum performance. In the first example we perform automatic drum hand recognition utilizing a technique for automatic labeling of training data using direct sensors, and only indirect sensors (e.g. a microphone) for testing. Left/Right drum hand recognition is achieved with an average accuracy of 84.95% for two performers. Secondly we provide a study investigating multimodality dependent performance metrics analysis.

Keywords

Multimodality, Drum stroke identification, surrogate sensors, surrogate data training, machine learning, music information retrieval, performance metrics

1. INTRODUCTION AND MOTIVATION

Combining machine learning techniques with percussive musical interface/instrument design is an emerging area of research that has seen many applications in recent years. Tindale investigated drum timbre recognition [17] and later applied similar techniques to turn regular drum triggers into expressive controllers for physical models [16]. Other examples have been proposed which even enable human-machine interaction with mechanical percussionists who can listen to human performers and improvise in real time [19].

In terms of signal processing there is now robust onset detection algorithms [1, 4] enabling one to accurately identify when musical events occur. Researchers have also been actively investigating other areas of musical performance such as tempo estimation [2, 7], beat tracking [3, 9], and percussive instrument segmentation [8]. Combining many of these techniques together, researches have explored the task of automatic transcription of drum and percussive performance, [6, 7, 12, 18]. While great advances have been made in the aforementioned tasks the majority of research into drum interaction scenarios which combine musical interfaces/instruments and machine learning have been concerned with the segmentation or isolation of individual drums from a recorded audio signal. While mono- and polyphonic drum segmentation is a major aspect to tasks such as automatic drum transcription, a key feature of drum performance (that has yet to be explored in current drum analysis literature) pertains to the physiological space of drum performance. Not only is it important to know when and which drum is played in a pattern, but also *which hand* is striking the drum. In this research we investigate this question, and propose a multimodal signal processing system for the automatic labeling and classification of left and right hand drum strikes from a monophonic audio source.

There are many real-world cases where drum stroke recognition is important. In fact most traditional exercises which practicing drummers study emphasize the practice of specific left and right hand patterns (more information on this can be found in section 2.1.3). In automatic transcription scenarios, a key element that has been missing up until now is transcribing which hand performed a particular drum hit. In order to fully understand ones performance it is important to know how the player moves around the drum(s), the nuances and differences present in the strikes of their independent hands, and the possible stylistic signifiers resulting from the physical aspects of their individual hand strikes. This presents a large problem, as it is nearly impossible to determine which hand is hitting a drum from a monophonic audio recording alone.

Using direct sensors such as accelerometers on the performers hands however we can capture extremely accurate ground truth about the movements of the performers hands. This comes at the cost of being invasive and possibly hindering performance. In a typical controlled machine-learning situation we can of course place constraints on the data-capturing scenario. One solution would be to only record left hand strikes, and then separately record right hand strikes, labeling them accordingly when performing feature extraction. We are interested however not only capturing each hand playing in isolation, but in context of actual performance and practice scenarios; and so the interplay between left and right hand playing is of utmost importance. Another option would be to manually label each audio event as being either from the left or right hand, based on a priori knowledge of a specific pattern played. As many data capturing scenarios (including ones in the research) involve specific patterns to be played, this is a common but time-consuming approach to labeling drum training performance data. Additionally this task is non-sympathetic to inevitable playing mistakes in the performance, which require manual adjustment when labeling the training data. We are also interested in investigating the improvisatory elements of drum performance, making the task of manually labeling hand-patterns nearly impossible. To overcome these challenges, this research turns to an exciting new technique inspired from Surrogate Sensing [14] to enable the automatic labeling of drum hand patterns for classification.

NIME'12, May 21-23, 2012, University of Michigan, Ann Arbor.

One of the earliest studies of drum performance showed how physical factors such as the start height of a stick could impact the resulting amplitudes and durations of the sound produced [10]. More recently, Dahl showed similar relationships between the correlation of strike velocity and the height and shape of the stroke in user studies [2]. Dolhansky et al. modeled the shape of a percussive stroke to turn mobile phones with accelerometers into physically-inspired percussive instruments [5]. There are many ways which people have attempted to analyze the gesture of drum performance and its effect on the dynamic and timbre spaces; Tindale et al. provides a good overview of sensor capturing methodologies in [15]. The research mentioned and other countless examples confirm the strong link between the physical space in which a performers actions exist, and the fingerprint imparted on the musical output. To this end we begin to investigate these ties in this paper by not only looking at drum-hand recognition, but also at statistical measures afforded by multimodal analysis of acoustical instrument output paired with NIME's.

The remainder of this paper is as follows: In section 2 we provide an overview of our data collection and analysis framework, including our implementation of surrogate data training for automatic hand labeling of training data. We show our drum hand recognition results in section 3 and performance metrics in section 4. Finally our conclusions are discussed in section in section 5.

2. SYSTEM DESIGN AND IMPLEMENTATION

In this section we describe the data capturing and analysis system used in the drum-stroke recognition experiment. From a high-level view, the drum-hand recognition experiment employs a three-step process including a data collection phase, an analysis phase, and finally the testing and machine-learning phase as illustrated in Figure 1.

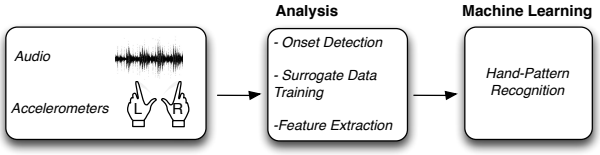

Figure 1 – Overview of Drum Hand Recognition System

2.1 Data Collection

A primary goal in the research was to make sure that the techniques used could easily be used in a variety of scenarios from live performance to assisted learning in music schools. It was also a goal to empower musicians to be able to run the experiments themselves. Other methodologies were considered, including hi-speed video camera tracking. While hi-speed video tracking could be a useful solution for hand tracking, and has been used by others for similar tasks such as bow performance tracking [13, 21], we desired a solution that was more affordable than typical hi-speed cameras, and that needed little to no calibration. Additionally the research was concerned with investigating surrogate data training, and so the following software and sensor system was used.

2.1.1 Nuance

Nuance is a program written by the authors as a general-purpose multi-track recording solution for multimodal data sources. Geared towards machine learning and musical data mining, Nuance enables nearly any musical sensor system and instrument communicating over serial, MIDI and/or OSC to synchronously capture its data to disk in .wav audio format. Once recorded the data is ready to be analyzed in other platforms such as Matlab, Marsyas, ChucK, etc. In our experiments all data was recorded as uncompressed .wav files at a sampling rate of 44100 samples per second. As audio-rate sampling is typically higher than common instrument sensor systems, sensor data is up-sampled to audio-rate using a sample-and-hold step function driven by the audio clock. This ensures synchronicity between audio and sensor data and enables the treatment of the sensor data as normal audio during analysis. Interacting with Nuance follows a similar paradigm as common multi-track digital audio workstations, and provides and easy drag-and-drop user interface.

2.1.2 Sensor System

In our experiments Nuance was used to synchronously record three axis of motion from two accelerometers placed on the hands of the performers, as well as a single mono microphone recording the acoustic drum signal. The ADXL335 tri-axis accelerometer was used, as well as a Shure SM57 for recording the audio output of the snare drum. The microphone was placed in a typical configuration, approximately 1" – 3" from the rim of the snare, slightly angled down and across the head of the drum.

The two accelerometers were placed on the topsides of the performers hands, and connected to a wireless transmitter. An Arduino Fio inside the transmitting device receives data form each axis of the accelerometers with 10-bit resolution, and transmits each sensor readings to a nearby computer over wireless Xbee (ZigBee) RF communication. This data was recorded directly over a serial-connection with the receiving Xbee module using Nuance.

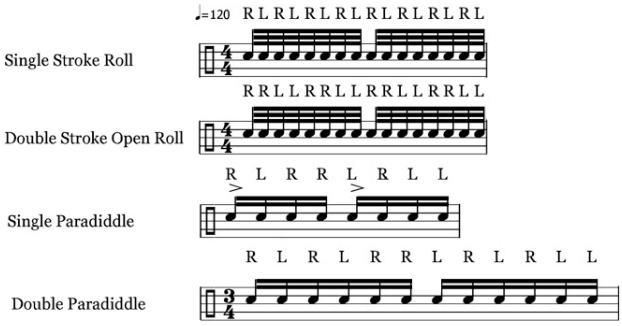

Figure 2 – Drum Rudiments Performed

2.1.3 Data Set

The system described in sections 2.1.1 and 2.1.2 was used to record a total of 2917 snare-drum hits from two performers. Performer one was at a beginner level whereas performer two was an intermediate/advanced percussionist. The performers were instructed to play four fundamental drum exercises from the Percussive Arts Society[1] International Drum Rudiments; these included the Single Stroke Roll (referred to as D1 throughout the remainder of the paper), the Double Stroke Open Roll (D2), the Single Paradiddle (D3), and the Double Paradiddle (D4) (Figure 2). Each exercise was recorded for roughly 3 minutes, resulting in a total of 1467 hits (736 right hand / 731 left hand) for performer one and 1450 hits (726 right hand / 724 left hand) for performer two. In preliminary testing the performers recorded purely improvisational, however a more regimented routine was played during the final data

[1] The PAS is the world's largest international percussion organization. More information on the PAS can be found at http://www.pas.org/

collection process to enable other research into specific performance metrics and rudiment classification. Figure 2 details the drum rudiments performed.

2.2 Analysis Framework

In the following sections we discuss the analysis framework that was used to extract features for the left/right hand classification experiments and metrics tracking.

2.2.1 Surrogate Data Training

One of the biggest hurdles for musical supervised machine learning is obtaining and labeling a large enough training data set for true results. As described earlier in section 1 manually labeling the training data is not an efficient process, nor does it easily deal with errors that are common in the data collection phase. By using a process that can automatically label training data, the training regiment can be more loosely defined, even allowing the performer to improvise (unless there was specific desire to record particular patterns as in our case). Common disturbances in the data collection process such as performance mistakes, which normally must be accounted for by the researchers manually are also no longer an issue. We turn to a new technique inspired by Surrogate Sensors [14] enabling us to quickly record and label each hit in our audio recordings by using known information from direct sensors (accelerometers) to navigate unknown information in the data from our indirect sensor (microphone). The direct sensors provide the benefit of near perfect ground truth making the technique extremely robust (see section 2.2.3). The method is also transferable to other sensors and modalities, and the particular implementation in this research is described in the following section on onset detection.

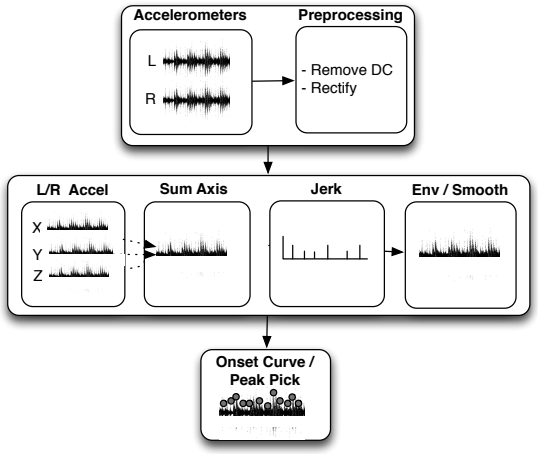

Figure 3 – Overview of Onset Detection Algorithm

2.2.2 Onset Detection

A triple-axis accelerometer was placed on each of the performers hands while recording the data sets. The ultimate goal was to use gesture onsets in the independent hands accelerometers to navigate and label the note onsets in the audio recordings. As shown in Figure 3, each axis (per accelerometer) is first preprocessed in Matlab by removing the DC offset and full-wave rectification. The accelerometers each have their three axis summed and averaged to collapse the data streams into a single dimension. Next jerk is calculated for each accelerometer, followed by a threshold function to remove spurious jitter. To further smooth the signals before onset detection is applied, the envelopes of the signals are extracted, and smoothed with a Gaussian of standard deviation of samples. The onset curve is then calculated and peak-picked at local maxima's. Lastly onset detection was also performed on the audio recording, and all three streams' (1 audio, 2 accelerometer) onset locations (in seconds) are stored in independent vectors. More detailed information on the onset detection algorithm can be found in [11].

2.2.3 Onset Detection Accuracy

Table 1 shows the onset classification accuracy of the accelerometers prior to correction. The high yield (99%) in accuracy of the accelerometers makes them a great candidate for surrogate labeling the audio onsets as either left or right hand onsets. The onset vectors were also exported as .txt files and imported into a beat-tracking application called BeatRoot [3] to visualize and (manually) correct any errors in the accelerometer onsets detected. It should be noted that the correction step was not necessary as the minor amount falsely detected onsets were few enough to not impact the data too much, however we desired 100% ground truth and so any false-positive and false-negative onsets were corrected in BeatRoot prior to feature extraction.

	Precision (L/R)	**Recall (L/R)**	**F-Measure (L/R)**
Performer 1	0.997	0.997	0.997
Performer 2	0.993	0.997	0.995

Table 1 – Accelerometer Onset Detection Accuracy

2.2.4 Feature Extraction

After onset detection, features were extracted in Matlab by taking the accelerometer onset positions for each hand and searching for the nearest detected onset (within a certain threshold determined by the frequency and tempo of the strikes) in the audio onsets. The strike in the audio file is then windowed to contain the entire single-hit and various features are extracted. The feature vector is labeled with the appropriate class (1 = Right, 2 = Left) and exported as an .arff file for machine learning analysis in Weka [20]. For each strike a 14-dimension feature vector is calculated containing: RMS, Spectral Rolloff, Spectral Centroid, Brightness, Regularity, Roughness, Skewness, Kurtosis, Spread, Attack Slope, Attack Time, Zero Crossing, MFCC (0^{th} coeff.), and the Onset Difference Time (ODT) between the detected audio and corresponding accelerometer onsets.

3. DRUM HAND RECOGNITION

After the data was collected it was imported into Weka for supervised learning. The primary focus of this experiment was to investigate if a machine could be trained to reliably classify which hand was used to strike a snare drum.

3.1 Classification

Five classifiers were used in our tests including a *Multilayer Perception* back-propagation artificial neural network, the *J48* decision tree classifier, *Naive Bays*, a support vector machine trained using *Sequential Minimal Optimization (SMO)*, and Logistic Regression. 10-Fold cross validation was used in all tests with a 12-dimension feature subset (attack time and onset difference features were removed for this experiment).

3.2 Results and Discussion

This section describes the outcomes obtained from our classification tests. As this is binary classification scenario (classification can either be left or right hand), the chance classification baseline is 50%.

Using the entire data set and 10-fold cross validation, the best results were achieved using multilayer perceptron (MLP) for

both performers. MLP yielded an accuracy of 84.93% for performer one and 84.96% classification accuracy for performer two. All of the algorithms appear to do a decent job at generalizing over the entire data set and provide similar classification results with smaller subsets of the feature vector.

	Performer 1 (%)	Performer 2 (%)
MLP	*84.93*	*84.96*
SMO	81.66	80.34
Naive Bays	75.05	63.65
Logistic	84.38	83.51
J48	81.93	78.62

Table 2 - Classification Accuracy Using All Data

While it is clear that some classifiers seem to generalize quite well in all cases, more simple probabilistic classifiers such as Naive Bays seem to benefit greatly from having a larger training set that covers a wider variance in feature data.

4. PERFORMANCE METRICS

Automatic drum hand recognition proposes exciting new possibilities including: more nuanced automatic drum transcription, preservation of performance technique from master musicians long after life, providing new controller data for live performance, and providing insightful information and metrics during regimented practice and musical training. However, the information from direct sensors can also be used in conjunction with indirect sensors to provide insightful new performance metrics and features. In this section we will look at new features and how they may add to our ability in describing and deducing meaningful information from musical performance.

4.1 Onset Differences

In traditional drum performance analysis, temporal information such as timing deviations and onsets of drum hits are normally investigated by analyzing an audio recording. Researchers have not only investigated the physical onset times (in audio) but have also looked at the perceptual onset and attack times (often called PAT) in order to measure when sounds are actually heard [5]. Here we consider the physical onset times from sensors on the actual performer in relation to the onset times recorded simultaneously in the acoustical output in what we call the Onset Difference Time or ODT.

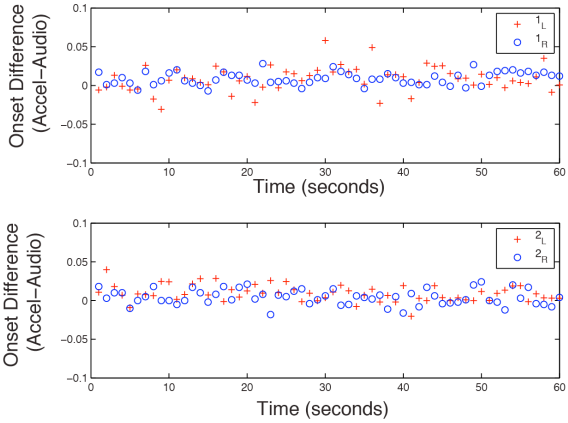

Figure 4 - Onset Difference Times for the first 60 sec. of D1 (Performer 1 Top, Performer 2 Bottom)

In Figure 4 we can see the onset difference times (in seconds) between the left (+) and right (o) hand accelerometer onsets and their audio onset times. A horizontal line at 0 would mean a perfect match (zero difference) in onset times, and an observation of the graphs shows that performer two (bottom) had a generally lower onset differentiation than performer one (top). Performer two was in fact a more highly experienced drummer, suggesting a great link or consistency in physical vs. acoustical onsets in this particular exercise. Observing Figure 4 it is also apparent that in this 60-second pass of D1, the onset difference times of performer two's individual hands were more closely related (in terms of mean onset difference) than that of performer ones.

Data Set	Min (rush)	Max (lag)	Mean	Std
P1	-0.0076	0.0148	0.0118	0.0116
P2	-0.0070	0.0149	0.0107	0.0133

Table 3 – Average Onset Difference Statistics for Both Performers

Table 3 and Figure 5 show averages from both hands and all data sets D1-D4. Min which we call "rush" is calculated as the average amount the accelerometer onsets that were earlier than the audio onsets. As such it is calculated only over negative onset difference times. On average, when performer two's physical strike onsets rushed the audio onsets, it did so less drastically than performer one. Again this may be attributed to the fact that performer two was a more experienced player with tighter timing than performer one.

Max or lag is calculated as the average amount the accelerometer onsets were later than their paired audio onsets (positive difference times). Coincidentally, both performers lag differences were extremely similar.

Mean is the average onset difference time calculated over the entire vector or onset differences for each performer. Again performer two performed with less distance between physical and audio onset times. Interestingly, performer two's standard deviation was slightly larger than performer one, meaning that the amount of dispersion from the performers mean performance was greater.

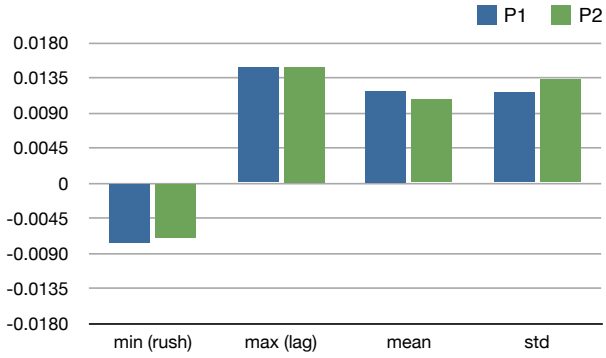

Figure 5 – Bar Graph Visualizing Table 4 Metrics

5. CONCLUSIONS AND FUTURE WORK

In this paper we investigated two ways in which multimodal signal processing and sensor systems can benefit percussive computation. In the first case study we used direct sensors (accelerometers) on a performer to automatically annotate and train the computer to perform drum stroke recognition from indirect sensors (a single microphone). Averaging the best results from two performers multilayer perception achieved 84.95% accuracy and shows that it is possible for the computer to identify whether a performer hit a drum with their left or right hand. Once trained with the direct sensors, the computer can non-invasively transcribe the physical attributes of a

percussionist's performance, adding important nuance to future automatic music transcription. Additionally automatic drum-hand recognition will be useful in many pedagogical scenarios such as rudiment identification, accuracy and other performance/metrics measures. In live performance contexts where it may be desired to trigger musical events, processes, and/or visualization based on particular sequences of strikes, drum stroke recognition using non-invasive methods will also be extremely powerful.

In the second case study we looked at statistics measures as performance metrics obtainable using a multimodal system. Our preliminary findings comparing data from typical direct and indirect sensors such as accelerometers and microphones (respectively) reconfirm the importance of the looking at both the audio-space and physical-space (simultaneously) when investigating musical performance. Research often chooses one or the other for analysis, however investigating the space in between is one we are excited at looking at more closely in the future.

In the future we are looking forward to expanding the data set with a larger pool of performers, as well as investigating how well the techniques generalize to different snare drums (and eventually other drums in the drum set). It would also be particularly useful to add a third strike to the test set, when a player performs more complex patterns, including striking with both hands. The authors are particular interested in performance metrics tracking, and so the techniques discussed in this paper will serve as a foundation to continue research into performance metrics tracking, allowing performers to evaluate their playing in live performance and in the practice room.

At the core of much of this is the trade-off between direct and indirect sensors. Indirect sensors such as microphones have proven to be extremely useful and reliable sources for music information retrieval, with the benefit on not hindering performance. At the same time they lack certain physical attributes that are only possible to obtain by placing more invasive direct sensors on the performer, and/or instrument/NIME. In one sense we hope this research brings wider attention to a part of a novel technique called surrogate sensing which reduces the negative impact of invasive sensors by constraining their dependency to the training phase of a musical system or experiment. At the same time there is lots of work ahead and so the future will definitely still hold an important space for direct sensors in these scenarios. We look forward to a future where direct sensors such as accelerometers are small and light enough to be embedded within a drum stick without altering performance in any way, but also one where a trained machine can play back a recording from great musicians of the past and automatically transcribe the magical nuances of their performances for future generations.

6. ACKNOWLEDGMENTS

We'd like to thank Adam Tindale and George Tzanetakis for their previous work in drum performance analysis and surrogate sensing.

7. REFERENCES

[1] Bello, J. et al. 2005. A Tutorial on Onset Detection in Music Signals. *Speech and Audio Processing, IEEE Transactions on*. 13, 5 (2005), 1035–1047.

[2] Dahl, S. 2005. *On the Beat: Human movement and Timing in the Production an Perception of Music*. Royal Institute of Technology.

[3] Dixon, S. 2007. Evaluation of the Audio Beat Tracking System BeatRoot. *Journal of New Music Research*. 36, 1 (2007), 39 – 50.

[4] Dixon, S. 2006. Onset Detection Revisited. *Proceedings of the 9th International Conference on Digital Audio Effects (DAFX'06)* (Montreal, Canada, 2006).

[5] Dolhansky, B. et al. 2011. Designing an Expressive Virtual Percussive Instrument. *Proceeding of the Sound and Music Computing Conference* (2011).

[6] Fitzgerald, D. 2004. Automatic Drum Transcription and Source Separation. *Doctoral*. (Jun. 2004).

[7] Gillet, O. and Richard, G. 2008. Transcription and Separation of Drum Signals From Polyphonic Music. *IEEE Transactions on Audio, Speech, and Language Processing*. 16, 3 (Mar. 2008), 529–540.

[8] Goto, M. and Muraoka, Y. 1994. A sound source separation system for percussion instruments. *Transactions of the Institute of Electronics, Information and Communication Engineers* (1994), 901–911.

[9] Goto, M. and Muraoka, Y. 1999. Real-time beat tracking for drumless audio signals: chord change detection for musical decisions. *Speech Commun*. 27, 3-4 (1999), 311–335.

[10] Henzie, C.A. 1960. *Amplitude and duration characteristics of snare drum tones*. Indiana University.

[11] Lartillot, Olivier et al. 2008. A Unifying Framework for Onset Detection, Tempo Estimation, and Pulse Clarity Prediction. *11th International Conference on Digital Audio Effects* (Espoo, Finland, Sep. 2008).

[12] Paulus, J.K. and Klapuri, A.P. 2003. Conventional and periodic N-grams in the transcription of drum sequences. *2003 International Conference on Multimedia and Expo, 2003. ICME '03. Proceedings* (Jul. 2003), II– 737–40 vol.2.

[13] Schoonderwaldt, E. et al. 2006. Combining accelerometer and video camera: reconstruction of bow velocity profiles. *Proceedings of the 2006 conference on New interfaces for musical expression* (Paris, France, France, 2006), 200–203.

[14] Tindale, A. et al. 2011. Training Surrogate Sensors in Musical Gesture Acquisition Systems. *IEEE Transactions on Multimedia*. 13, 1 (Feb. 2011), 50–59.

[15] Tindale, A.R. et al. 2005. A comparison of sensor strategies for capturing percussive gestures. *Proceedings of the 2005 conference on New interfaces for musical expression* (Singapore, Singapore, 2005), 200–203.

[16] Tindale, A.R. 2007. A hybrid method for extended percussive gesture. *Proceedings of the 7th international conference on New interfaces for musical expression* (New York, NY, USA, 2007), 392–393.

[17] Tindale, A.R. et al. 2004. Retrieval of percussion gestures using timbre classification techniques. *ISMIR* (2004).

[18] Tzanetakis, G. et al. 2005. Subband-based Drum Transcription for Audio Signals. (Oct. 2005), 1–4.

[19] Weinberg, G. and Driscoll, S. 2006. Robot-human interaction with an anthropomorphic percussionist. *Proceedings of the SIGCHI conference on Human Factors in computing systems* (New York, NY, USA, 2006), 1229–1232.

[20] Witten, I.H. et al. 2011. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. Morgan Kaufmann.

[21] Zhang, B. et al. 2007. Visual analysis of fingering for pedagogical violin transcription. *Proceedings of the 15th international conference on Multimedia* (New York, NY, USA, 2007), 521–524.