Learning Indirect Acquisition of Instrumental Gestures using Direct Sensors

(Invited Paper)

George Tzanetakis, Ajay Kapur, Adam R. Tindale
Department of Computer Science
University of Victoria
Victoria, British Columbia, CANADA
Email: gtzan@cs.uvic.ca, ajay@ece.uvic.ca, art@uvic.ca

***Abstract*— Sensing instrumental gestures is a common task in interactive electroacoustic music performances. The sensed gestures can then be mapped to sounds, synthesis algorithms, visuals etc. Two of the most common approaches for acquiring these gestures are 1) Hybrid Instruments which are "traditional" musical instruments enhanced with sensors that directly detect gestures 2) Indirect Acquisition in which the only measurement is the acoustic signal and signal processing techniques are used to acquire the gestures. Hybrid instruments require modification of existing instruments which is frequently undesirable. However they provide relatively straightforward and reliable measuring capability. On the other hand, indirect acquisition approaches typically require sophisticated signal processing and possibly machine learning algorithms in order to extract the relevant information from the audio signals. In this paper the idea of using direct sensors to train a machine learning model for indirect acquisition is explored. This approach has some nice advantages, mainly: 1) large amounts of training data can be collected with minimum effort 2) once the indirect acquisition system is trained no sensors or modifications to the playing instrument are required. Case studies described in paper include 1) strike position on a snare drum 2) strum direction on a sitar.**

I. Introduction

Throughout history musical instruments have been some of the best examples of artifacts designed for interaction. In recent years a combination of cheaper sensors, more powerful computers and rapid prototyping software have resulted in a plethora of interactive electroacoustic music performances and installations. In many of these performances, traditional acoustic instruments are blended with computer generated sounds and visuals. Automatically sensing gestures is frequently desired in such interactive multimedia performances.

There are two main approaches to sensing instrumental gestures. In direct acquisition, traditional acoustical instruments are extended/modified with a variety of sensors such as force sensing resistors (FSR), and accelerometers. The purpose of these sensors is to measure various aspects of the gestures of the performers interacting with their instruments. A variety of such "hyper" instruments have been proposed. However, there are many pitfalls in creating such sensor-based controller systems. Purchasing microcontrollers and certain sensors can be expensive. The massive tangle of wires interconnecting one unit to the next can get failure-prone. Things that can go wrong include: simple analog circuitry break down, or sensors wearing out right before a performance forcing musicians to carry a soldering iron along with their tuning fork. The biggest problem with hyper instruments, is that there usually is only one version, and the builder is the only one that can benefit from the data acquired and use the instrument in performance.

These problems have motivated researchers to work on indirect acquisition in which the musical instrument is not modified in any way. The only input is provided by non-invasive sensors typically one or more microphones. The recorded audio then needs to be analyzed in order to measure the various gestures. Probably the most common and familiar example of indirect acquisition is the use of automatic pitch detectors to turn monophonic acoustic instruments into MIDI (Music Instrument Digital Interface) instruments. In most cases indirect acquisition doesn't directly capture the intended measurement and the signal needs to be analyzed to extract the information. In most cases this analysis is achieved by using real-time signal processing techniques. More recently an additional stage of supervised machine learning has been utilized in order to "train" the information extraction algorithm. The disadvantage of indirect acquisition is the significant effort required to develop the signal processing algorithms. In addition, if machine learning is utilized the training of the system can be time consuming and labor intensive.

The main problem addressed in this paper is the efficient and effective construction of indirect acquisition systems for musical instruments in the context of interactive media. Our proposed solution is based on the idea of using direct sensors to train machine learning models that predict the direct sensor outputs from acoustical data. Once these indirect models have been trained and evaluated, they can be used as "virtual" sensors in place of the direct sensors. This approach is motivated by ideas in multimodal data fusion with the slight twist that in our case the data fusion is only used during the learning phase. We believe that the idea of using direct sensors to learn indirect acquisition can be applied to other area of multimodal interaction in addition to musical instruments.

This approach of using direct sensors to "learn" indirect acquisition models has some nice characteristics. Large amounts of training data can be collected with minimum effort just by playing the enhanced instrument with the sensors. Once the system is trained and provided the accuracy and performance

of the learned "virtual" sensor is satisfactory there is no need for direct sensors or modifications to the instrument.

The traditional use of machine learning in audio analysis has been in classification where the output of the system an ordinal value (for example the instrument name). As a first case study of our proposed method we describe a system for classifying percussive gestures using indirect acquisition. More specifically the strike position of a stick on a snare drum is automatically inferred from the audio recording. A radio drum controller is used as the direct sensor in order to train the indirect acquisition. In addition, we explore regression which refers to machine learning systems where the output is a continuous variable. One of the challenges in regression is obtaining large amounts of data for training which is much easier using our proposed approach. In our experiments, we use audio-based feature extraction with synchronized continuous sensor data to train a "virtual" sensor using machine learning. More specifically we describe experiments using the Electronic Sitar (ESitar), a digitally enhanced sensor based controller modeled after the traditional North Indian sitar.

II. Background

The use of sensors to gather gestural data from a musician has been used as an aid in the creation of real-time computer music performance. In the last few years the New Interfaces for Musical Expression (NIME) conference has been the main forum for advances in that area. Some representative examples of such systems are: the Hypercello [1], the digitized Japanese drum Aobachi [2], and the E-Sitar [3]. All These instruments still function as acoustical instruments but are enhanced with a variety of direct sensors to capture the gestures.

In addition, there has been some research using machine learning techniques [4] to classify specific gestures based on audio feature analysis. The extraction of control features from the timbre space of the clarinet is explored in [5]. Deriving gesture data from acoustic analysis of a guitar performance is explored in [6]–[8]. An important influence for our research is the concept of indirect acquisition of instrumental gesture described in [8]. Gesture extraction from drums is explored in [9]–[11]. The proposed algorithms rely on signal processing possibly followed by machine learning to extract information. Typically the information is categorical in nature for example the type of drum sound played (for example snare, bass drum or cymbal). In such approaches a large number of drum sounds are collected, labeled manually, and then used with audio feature extraction to train machine learning models.

In this paper, we address the challenge of collecting large amounts of training data without needing to manually label recordings. Direct sensors are used to automatically annotate the recordings. Once the indirect acquisition method has achieved satisfactory performance the direct sensors can be discarded. Collecting large amounts of data becomes simply playing the instrument. Most existing indirect acquisition methods make categorical decisions (classification). Using regression [12] it is possible to deal with continuous gestural data in a machine learning framework. However training regression models requires more data which is much easier using the proposed approach rather than manual labeling.

III. Audio Analysis

A. Audio-Based Feature Extraction

The feature set used in this paper is based on standard features used in isolated tone musical instrument classification, music and audio recognition. For the E-Sitar experiments it consists of 4 features computed based on the Short Time Fourier Transform (STFT) magnitude of the incoming audio signal. It consists of the Spectral Centroid (defined as the first moment of the magnitude spectrum), Rolloff and Flux as well as RMS energy. More details about these features can be found in [13]. The features are calculated using a short time analysis window with duration 10-40 milliseconds. In addition, the means and variances of the features over a larger texture window (0.2-1.0 seconds) are computed resulting in a feature set with 8 dimensions. The large texture window captures the dynamic nature of spectral information over time and it was a necessary addition to achieve any results in mapping features to gestures. Ideally the size of the analysis and texture windows should correspond as closely as possible to the nature time resolution of the gesture we want to map. In our experiments we have looked at how these parameters affect the desired output. In addition, the range of values we explored was determined empirically by inspecting the data acquired by the sensors. For the snare drum experiments the analysis window is 40 msecs (no texture window) and the features used are Centroid, Rolloff well as Mel-frequency Cepstral Coefficients (MFCCs). A preprocessing step of silence removal and onset detection ensure that features are only calculated once for each drum hit. The analysis window is located so that it captures most of the energy of the hit. The Marsyas [1] audio analysis and synthesis framework is used for the feature extraction and direct sensor acquisition and alignment with the audio features.

B. Regression and Classification

Classification refers to the prediction of discrete categorical outputs from real-valued inputs. A variety of classifiers have been proposed in the machine learning literature [4] with different characteristics in respect to training speed, generalization, accuracy and complexity. The main goal of the paper is to provide evidence to support the idea of using direct sensors to train models. Therefore experimental results are provided using a few representative classification methods.

Regression refers to the prediction of real-valued outputs from real-valued inputs. Multivariate regression refers to predicting a single real-valued output from multiple real-valued inputs. A classic example is predicting the height of a person using their measure weight and age. There are a variety of methods proposed in the machine learning [4] literature for regression. For the experiments described in this paper, we use linear regression where the output is formed as a linear combination of the inputs with an additional constant factor.

[1] `http://marsyas.sourceforge.net`

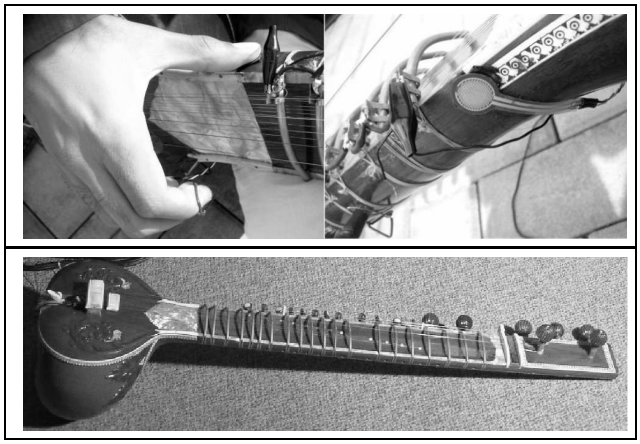

Fig. 1. E-Sitar and thumb sensor

TABLE I

EFFECT OF ANALYSIS WINDOW SIZE.

Analysis Window Size	128	256	512
Correlation Coefficient	0.2795	0.3226	0.2414

TABLE II

EFFECT OF TEXTURE WINDOW SIZE (COLUMNS) AND REGRESSION METHOD (ROWS).

	10	**20**	**30**	**40**
Random Output	0.14	0.14	0.14	0.14
Linear Regression	0.28	0.33	0.28	0.27
Neural Network	0.27	0.45	0.37	0.43
M5' Regression Method	0.28	0.39	0.37	0.37

Linear regression is fast to compute and therefore useful for doing repetitive experiments for exploring the parameter. We also employ a more powerful back propagation neural network [4] that can deal with non-linear combinations of the input data. The neural network is slower to train but provides better regression performance. Finally, the M5 prime decision tree based regression algorithm was also used. The performance of regression is measured by a correlation coefficient which ranges from 0.0 to 1.0 where 1.0 indicates a perfect fit. In the case of gestural control, there is significant amount of noise and the sensor data doesn't necessarily reflect directly the gesture to be captured. Therefore, the correlation coefficient can mainly be used as a relative performance measure between different algorithms rather than an absolute indication of audio-based gestural capturing. The automatically annotated features and direct sensor labels are exported into the Weka [2] machine learning framework for training and evaluation.

IV. CASE STUDIES

A. ESitar

The ESitar was built with the goal of capturing a variety of gestural input data. A more detailed description of audio-based gesture extraction on the ESitar including monophonic pitch detection can be found in [12]. A variety of different sensors such as fret detection using a network of resistors are used combined with an Atmel AVR ATMega16 microcontroller for data acquisition. The data is sent out using the MIDI protocol. In this paper we describe how to indirectly acquire the *mizrab* pluck direction.

On the right index finger, a sitar player wears a ring like plectrum, known as the *mizrab*. The right thumb, remains securely on the edge of the *dand* (neck) as shown in Figure 1, as the entire right hand gets pulled up and down over the main seven strings, letting the mizrab strum. An upward stroke is known as *Dha* and a downward stroke is known as *Ra*.

The direct sensor used to deduce the direction of a *mizrab* stroke is a force sensing resistor (FSR), which is placed directly under the right hand thumb, as shown in Figure 1. The thumb never moves from this position while playing, however the applied force varies based on the *mizrab* stroke direction. A *Dha* stroke (upward stroke) produces more pressure on the thumb than a *Ra* stroke (downward stroke). We send a continuous stream of data from the FSR via MIDI, because this data is rhythmically in time and can be used compositionally for more than just deducing pluck direction. A vector of audio features is extracted and the values of the FSR sensor are fused and used to train the "virtual" sensor using a regression model.

Our first experiment was to analyze the effect of the analysis window size used for audio feature extraction. Table I shows the results. The texture size remained constant at 0.5 seconds and linear regression was used. The correlation coefficient for random inputs is 0.14. It is apparent based on the table that an analysis window of length 256 (which corresponds to 10 milliseconds) achieves the best results. It can also be seen that the results are significantly better than chance. We used this window size for the next experiment.

The next experiment explores the effect of texture window size and choice of regression method. Table II shows the results. The rows correspond to regression methods and the columns correspond to texture window sizes expressed in number of analysis windows. For example, 40 corresponds to 40 windows of 256 samples at 22050 Hz sampling rate which is approximately 0.5 seconds. To avoid overfitting we use a percentage split where the first 50% of the audio and gesture data recording is used to train the regression algorithm which is then used to predict the second half of recorded data.

B. Snare Drum

The snare drum is a modern drum common to most drum sets. Wire snares are attached to the bottom drumhead so that when the drum is struck on the top the snares vibrate against the drum, creating a distinctive timbre. Using a mechanism it is possible to disengage the snares and stop their vibration allowing the drum to have a more traditional timbre.

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

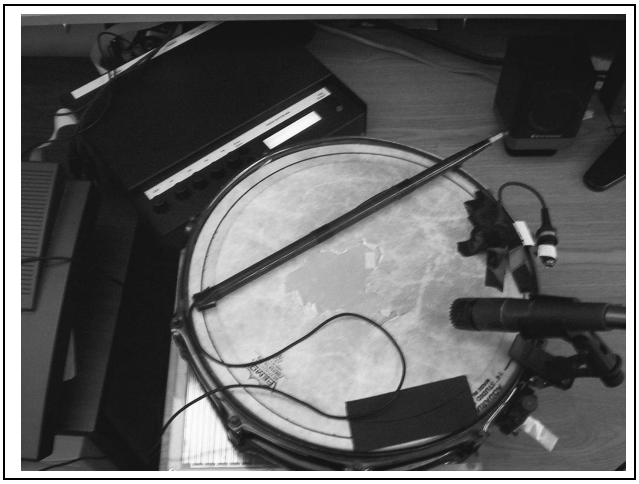

Fig. 2. Snare drum, radio drum stick, and microphone

TABLE III
PERCENTAGES OF CORRECTLY CLASSIFIED SNARE DRUM HITS

	ZeroR	NB	MLP	MLR	SMO
Snares	53	92	91	91	92
No Snares	57	93	94	95	95
Improvisation	59	79	77	78	78

The third author completed a Masters thesis [14] on the topic of indirect acquisition of snare drum gestures. In this thesis, 1260 samples were collected with three drums and three expert players. The process of collecting and processing the training data took nearly a week of manual labor. Using the method described in this paper the process took under an hour. The direct sensor used for training is the Radio Drum [15] which is based on capacitance sensors. It can detect the x,y,z positions of two drum sticks in 3D space. This allowed us to place the surface of the *Radio Drum* under a snare drum and still be able to measure the stick position. For each hit the radial position was measured and the hit was labeled as either "edge" or "center" using thresholding. Audio features are also extracted in real-time using input from a microphone. The features and sensor measurements are then used for training classifiers. The setup can be viewed in Figure 2.

Table III shows classification results using a variety of classifiers. The **Snares, No Snares** rows are calculated using approximately 1000 drum hits with the snares engaged/not engaged. All the results are based on 10-fold cross-validation. The trivial *ZeroR* classifier is used as a baseline. The following classifiers are used: *Naive Bayes* (NB), *Multi-Layer Perceptron* (MPL), *Multinomial Logistic Regression* (MLR), and *Support Vector* trained using sequential minimal optimization (SMO). The results are consistent between different classifier types and show that indirect acquisition using audio-based features trained using direct sensors is feasible. The **Improvisation** row is calculated using 200 drum hits of an improvisation. Even though the results are not as good as the cleaner previous rows they demonstrate that any performance can potentially be used as training data. A classically trained percussionist was used for data collection and no pre-processing or post-processing the classification results was performed.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the use of direct sensors to "train" machine learning model based on audio feature extraction for indirect acquisition. Once the model is trained and its performance is satisfactory the direct sensors can be discarded. That way large amounts of training data for machine learning can be collected with minimum effort just by playing the instrument. In addition, the learned indirect acquisition method allows capturing of non-trivial gestures without modifications to the instrument. We believe that the idea of using direct sensors to train indirect acquisition methods can be applied to other area of interactive media and data fusion.

There are many directions for future work. We are exploring the use of additional audio-based features Linear Prediction Coefficients (LPC) and sinusoidal analysis. We are also planning more extensive experiments with more instruments, players and desired gestures. Creating tools for further processing the gesture data to reduce the noise and outliers is another direction for future research. Another eventual goal is to use these techniques for transcription of music performances.

REFERENCES

[1] T. Machover, "Hyperinstruments: A progress report," MIT, Tech. Rep., 1992.

[2] D. Young and I. Fujinaga, "Aobachi: A new interface for japanese drumming," in *Proc. New Interfaces for Musical Expression (NIME)*, Hamamatsu, Japan, 2004.

[3] A. Kapur, P. Davidson, P. Cook, P. Driessen, and A. Schloss, "Digitizing north indian performance," in *Proc. Inter. Computer Music Conf. (ICMC)*, Miami, Florida, 2004.

[4] T. Mitchell, *Machine Learning*. Columbus, OH: McGraw Hill, 1997.

[5] E. B. Egozy, "Deriving musical control features from a real-time timbre analysis of the clarinet," Master's thesis, Massachusetts Institute of Technology, 1995.

[6] N. Orio, "The timbre space of the classical guitar and its relationship with plucking techniques," *Int. Computer Music Conf. (ICMC)*, 1999.

[7] C. Traube and J. O. Smith, "Estimating the plucking point on a guitar string," *Conference on Digital Auido Effects*, 2000.

[8] C. Traube, P. Depalle, and M. Wanderley, "Indirect acquisition of instrumental gestures based on signal, physical and perceptual information." *Proceedings of the Conference on New Musical Interfaces for Musical Expression*, pp. 42–7, 2003.

[9] F. Gouyon and P. Herrera, "Exploration of techniques for automatic labeling of audio drum tracks' instruments," *Proceedings of MOSART: Workshop on Current Directions in Computer Music*, p. [n.p.], 2001.

[10] J. Silpanpää, "Drum stroke recognition," Tampere University of Technology, Tampere, Finland, Tech. Rep., 2000. [Online]. Available: www.cs.tut.fi /sgn/arg/music/drums/raportti.ps

[11] A. Tindale, A. Kapur, G. Tzanetakis, and I. Fujinaga, "Retrieval of percussion gestures using timbre classifi cation techniques," *International Symposium on Music Information Retrieval*, 2004.

[12] A. Kapur, G. Tzanetakis, and P. F. Driessen, "Audio-based gesture extraction on the esitar controller," *Conference on Digital Auido Effects*, 2004.

[13] G. Tzanetakis and P. Cook, "Musical Genre Classifi cation of Audio Signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, July 2002.

[14] A. Tindale, "Classifi cation of snare drum sounds using neural networks," Master's thesis, McGill University, 2004.

[15] M. Mathews and W. Schloss, "The radio drum as a synthesizer controller," in *Proc. Int. Computer Music Conference (ICMC)*, Columbus, Ohio, 1989.