MULTIMODAL MUSICIAN RECOGNITION

Jordan Hochenbaum[1]
New Zealand School of Music[1]
PO Box 2332
Wellington, New Zealand
+64 (4) 463 5369
hochenjord@myvuw.ac.nz

Ajay Kapur[1,2]
California Institute of the Arts[2]
24700 McBean Parkway
Valencia, California, 91355
+1 (661) 952 3191
akapur@calarts.edu

Matthew Wright
Univ. California, Santa Barbara
CREATE/MAT, 2213 Elings Hall
Santa Barbara, California, 93106
+1 (805) 893 8352
matt@create.ucsb.edu

ABSTRACT

This research is an initial effort in showing how a multimodal approach can improve systems for gaining insight into a musician's practice and technique. Embedding a variety of sensors inside musical instruments and synchronously recording the sensors' data along with audio, we gather a database of gestural information from multiple performers, then use machine-learning techniques to recognize which musician is performing. Our multimodal approach (using both audio and sensor data) yields promising performer classification results, which we see as a first step in a larger effort to gain insight into musicians' practice and technique.

Keywords: Performer Recognition, Multimodal, HCI, Machine Learning, Hyperinstrument, eSitar

1. INTRODUCTION

We imagine a new multimodal language between musician and machine in which the computer receives multiple channels of information from the performer and interprets these data to derive information allowing for more effective communication back to the musician. It is important for the computer first to understand who is the performer, in order to tailor a specific and meaningful interaction. Moreover, we suggest that our musician recognition framework establishes a foundational multimodal language that can be the basis for future novel interactive and educational experiences between musicians and computers.

Two main approaches to performer recognition currently exist. The first uses audio-based techniques to identify characteristics from a recording [6-11, 13]. Stamatatos and Widmer explored this approach to quantify aspects of multiple players' performance "styles" and classify/identify performers using stylistic subtleties [8]. Their use of simple audio-based classifiers to distinguish among a small set of highly trained and stylistically polished players inspired our approach for data capturing.

The second approach is *multimodal*, combining audio with data from sensors capturing aspects of a performer's physical motion. Past research on other tasks in the field of Music Information Retrieval produced larger success rates through the use of multimodal instruments as compared to traditional audio-only approaches [2,3], while still maintaining transparency between user and instrument. An abundance of musical information resides not only in the sound produced, but also within the performer's physical interaction with the instrument, and we show that this physical information is beneficial to the difficult task of player identification.

To test our multimodal approach, we used a modified North Indian sitar [1] and performed player recognition over a group of 5 beginner, intermediate, and expert sitar players. The sitar is an extraordinarily difficult instrument to master, and requires very specific and demanding techniques for both the musician's left and right playing hands. Additionally, the instrument's character is rich an expressivity and allows each musician to develop an individual "style" of playing, adding individualized variability to the sitarist's technique. This makes the sitar a great candidate for empirical study of a particular player's technique, because the musical literature and tradition ask for specific physical actions to be performed by the musician, while the musician develops individual characteristics of his/her own.

The remainder of this paper is organized as follows. The Data Collection section describes our unique multimodal toolset of sensor-modified instruments and software that quantifies the physical input of performers. The Data Sets section describes the different musical material gathered for our various performer recognition experiments. System Design and Implementation describes the various features extracted from our data sets, and Results discusses the findings derived from the experiments.

2. DATA COLLECTION

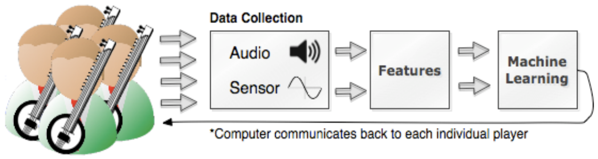


Figure 1 - Overview of the multimodal performer recognition system

This section discusses the various tools used for our experimentation. This toolset includes a custom built hyperinstrument [5], as well as a custom built software solution for capturing synchronous audio and sensor data. Figure 1 shows a general overview of our data capturing system. User(s) play a modified instrument (in our experiments, a sitar) and a computer captures the audio output and sensor data. The computer then extracts features from the performance and uses the features to perform player classification/recognition as discussed below. Lastly, the identification results are communicated back to the player.

2.1 ESitar

The ESitar is a multimodal hyperinstrument which has been transparently retrofitted with a unique sensor system [1]. Because the ESitar is a traditional sitar modified with sensors, it can be played and practiced as a regular sitar with very minimal adjustments. Not only does this make the transition between sitar and ESitar seamless for practiced players, but it also allows any sitar player (even beginners) to easily engage with the system, without having to become comfortable with an additional interface or instrument. For these experiments we use a force-sensing resistor (FSR) to capture the right hand (plucking) technique and a fret resistor network to identify the left hand (fretting) technique, as described in [1].

2.2 Super Recorder

Super Recorder is a software suite created to synchronously capture audio and sensor data from the ESitar. Super recorder has been designed as an additive framework, which easily allows for the integration of new sensor systems on both the ESitar and other hyperinstruments. The GUI for Super Recorder was programmed in the Processing programming environment, and communicates with the programming environment ChucK [12], via OSC [15]. The GUI provides an easy interface for setting up controlled experiments and visual feedback for all of the incoming data.

The current implementation of Super Recorder simultaneously collects an audio recording of the performance and data from the thumb FSR, the fret-sensing resistor network, and a three-axis accelerometer on the headstock of the ESitar, as well as from another three-way accelerometer placed on the performer's body to extract additional gestural information from the physical performance. The experiments described in this paper make use of only the audio recordings, thumb, and fret sensors, but we believe the accelerometer data will be useful in future research.

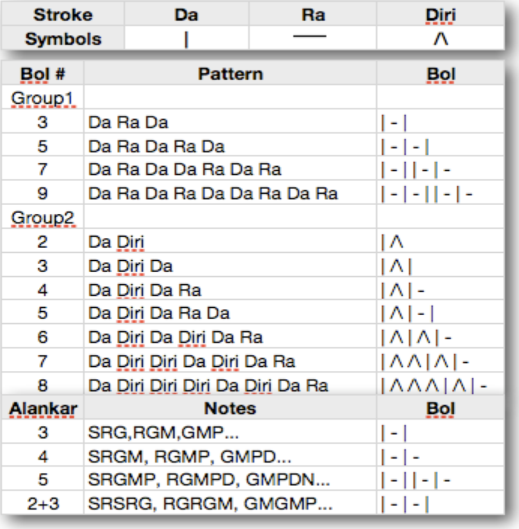
Data from all sensors are sampled at 100 Hz and stored as uncompressed wav files at a 44100 Hz sampling rate.[1] A metronome is also built into Super Recorder, allowing for more highly controlled and synchronous experiment set up.

3. Data Sets

We used the system described above to record three sitar performance data sets, ranging along a continuum from strictly codified material to improvisation. In each case we recorded audio, thumb, and fret sensor data from five musicians.

3.1 Data Set 1 – "Exercises" (Practice Routine)

Our first data set was designed to record a player's individual performance characteristics during disciplined practice exercises. We chose two central exercises from the vast literature of classical North Indian practice methods [4]: *Bol* patterns and *Alankars*. *Bol* patterns are specific patterns of *da* (up stroke), *ra* (down stroke), and *diri* (up stroke and then down stroke in rapid succession), which are explicitly used in sitar practice plucking training, as well as in performance.[2] *Alankars* refer to scalar patterns that can be modally transposed; they form the basis of many musical ornaments and are also often used for melodic development and fretting practice. We used the *Bol* patterns and *Alankar* exercises shown in Table 1, played in the Indian Rag *Yaman*[3] at 220 beats per minute. Each of these 15 exercises was repeated as necessary to achieve a duration of 60 seconds.



Stroke	Da	Ra	Diri
Symbols	|	—	Λ

Bol #	Pattern	Bol
Group1		
3	Da Ra Da	| - |
5	Da Ra Da Ra Da	| - | - |
7	Da Ra Da Da Ra Da Ra	| - | | - | -
9	Da Ra Da Ra Da Da Ra Da Ra	| - | - | | - | -
Group2		
2	Da Diri	| Λ
3	Da Diri Da	| Λ |
4	Da Diri Da Ra	| Λ | -
5	Da Diri Da Ra Da	| Λ | - |
6	Da Diri Da Diri Da Ra	| Λ | Λ | -
7	Da Diri Diri Da Diri Da Ra	| Λ Λ | Λ | -
8	Da Diri Diri Diri Da Diri Da Ra	| Λ Λ Λ | Λ | -
Alankar	**Notes**	**Bol**
3	SRG,RGM,GMP...	| - |
4	SRGM, RGMP, GMPD...	| - | -
5	SRGMP, RGMPD, GMPDN...	| - | | - | -
2+3	SRSRG, RGRGM, GMGMP...	| - | - |

Table 1 - *Bol* Pattern and *Alankar* Exercises (Data Set 1)

3.2 Data Set 2 – *"Yaman Gat"* (Composition)

A *gat* is a fixed instrumental composition that provides the main theme(s) of a performance. Data set 2 contains ten 60-second recordings of each performer (50 total) repeating a particular *gat* in *rag Yaman* [4] eight times at 132 bpm.

3.3 Data Set 3 – "Improv"

Data Set 3 consists of sensor data and recordings collected from five players each performing ten different 60-second long free improvisations. This data set is completely unconstrained in terms of performers' technique; it was designed to support experiments to determine whether player performance data is context/piece specific, or truly a technique-based identifier.

4. System Design and Implementation

We extracted features from the audio recordings, thumb, and fret sensors using Matlab, and then stored them in a feature matrix suitable for classification using the Weka Data Mining toolset.

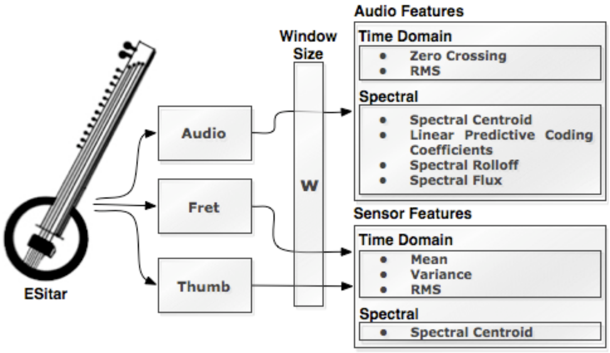


Figure 2 – Overview of Data Capturing and Feature Extraction Suite

4.1 Feature Extraction

Each sensor outputs continuous information and the recorded audio is also continuous, making the total amount of data a linear function of duration. For classification purposes, regardless of machine learning technique, we need a set of *features*, each of which collapses the recorded time-series data into fixed number of scalar quantities. We examined several

[1] Upsampling is with a simple step function: each sensor sample repeats 441 times in the output wav file.

[2] In general *da* represents the dominant stroke, which for sitar is upwards but for other instruments such as sarode is downwards.

[3] *Rag Yaman* uses the Lydian scale, i.e., major with a sharpened fourth scale degree.

features from both the audio and sensor data; Figure 2 shows the ones that yielded the best results).

4.1.1 *Thumb Pressure Features*

The arithmetic mean is a simple method of extracting a single characteristic average of the thumb pressure sensor data for each recording. Our player-pool included players from various skill-levels; we hypothesized that more highly trained sitarists might maintain a more consistent range of thumb pressure for the duration of a performance as compared to beginner players. To examine this hypothesis, variance was used. Spectral centroid was used to examine high-frequency transients produced while plucking, effectively relating to the amount of change (from subtle to jerky) in each player's plucking technique.

4.1.2 *Fret Features*

Fret *Mean*, *Variance, RMS,* and *Spectral Centroid* were also extracted from the fret sensor network. We were interested in determining if data from fretting tendencies and abilities could be effective player identifiers. For example, in data sets 1 and 2, fret mean could be an indicator of how frequently a player's left hand lost contact with the string. In data set 3, fret mean is a crude indicator of pitch register for each improvisation. The amount of fret variance per window could perhaps suggest the amount of distance and range covered by the players fretting hand at different moments of a performance.

4.2 Windowing

Each of our data sets consists of 60-second recordings. In addition to computing each feature once per 60-second performance, we also experimented with "windowing" the performances into non-overlapping time segments and computing the features once per segment. For example, with 10-second segments we divide each 60 second data recording into six 10-second "chunks" and compute our features for each chunk, multiplying our amount of training data by a factor of six by computing each feature over a smaller excerpt of music. (See the Windowing Results section below).

4.3 Classification

Five different classifiers were used in the machine learning experiments. These include a support vector machine trained using *Sequential Minimal Optimization* (SMO), a *multi-layer perceptron* (MLP) backpropagation artificial neural network*, IBk*, which implements the k-nearest-neighbors classifier, decision tree (*J48*), and *Naïve Bayes*. More detailed information about these classifiers and *Weka,* the data mining tool used in these experiements can be found in [14].

5. Results and Discussion

This section describes the outcomes obtained from our various machine-learning experiments. In each case we evaluated performance with 10-fold cross validation. Keep in mind that every data set had 5 performers, so chance performance would be 20% performer recognition accuracy.

5.1 Audio-only Results

This section demonstrates the classification results achieved by examining only the features extracted from the audio recordings. The advantage of this technique is that it can be performed with any instrumental player, using only the audio output of their instrument (either with a microphone or direct input), without requiring any modifications to the instrument.

Table 2 shows the classification results achieved using 3 different classifiers, for each data set alone, as well as all three data sets combined into one large corpus. Multilayer Perceptron proved to be the most accurate classifier in these tests, with the best accuracy being achieved on the exercises and *Yaman gat* composition data sets. For each pass in those two data sets, each player repeated the same sequence of defined notes/plucks for the duration of 60 seconds. Additionally, in data set 2, each pass contained the same pattern being played for its entirety. These two best accuracies may therefore be the result of slight data over fitting. Still, accuracy on the free improvisation data set, as well as combining all the data into one large pool yielded very satisfactory results.

	Exerc. (%)	***Yaman* (%)**	**Improv (%)**	**All (%)**
MLP	96.33	100	90	85
SMO	79.33	95.5	81	66.43
Naïve Bayes	87.33	98	71.5	58.14

Table 2 – Accuracy Achieved using Audio Only (15 Second Windowing)

5.2 Sensor-only Results

Table 3 shows the results of the same machine learning processes applied to only the sensor data. While the accuracy percentage achieved was slightly below the results from our audio features, the results are very exciting because they show that useful data does indeed reside in the musicians' physical/gestural information.

	Exerc. (%)	***Yaman* (%)**	**Improv (%)**	**All (%)**
MLP	84.33	100	89	75.15
SMO	63.67	100	67	60
Naïve Bayes	55.67	99	69	46

Table 3 - Accuracy Achieved using Sensors Only (15 Second Windowing)

Again, the highest accuracy was achieved using the *Yaman gat* data set, for which the sitarists were instructed to play the same scalar and plucking patterns repeatedly, for 10, 60 second long passes. We attribute part of the success of the achieved accuracy to the fact that the repetition asked of the players by the data set routine afforded the players ample time to get into a comfortable physical pattern, requiring the least amount of physical change and adjustment compared to the other data sets.

Using features derived only from our sensor data, the improvisation data set yielded the 2nd most accurate player identification across all three classifiers. While this could be the result of chance, it raises the possibility that when improvising, the musicians might have fallen into physical comfort-zones or patterns that they naturally tended to play. In the exercise routines (data set 1), for each pass the sitarists were required to change the fretting and plucking patterns to a hard defined set of practice routines. Because the exercise data set required specific plucking patterns that changed on each pass, and the improv data set allowed the musicians to freely play whatever came to them naturally, it is highly possible that specific plucking tendencies of the players technique were exposed through the improv data set, resulting in a higher classification accuracy than the exercise data set.

5.3 Single Sensor Feature Results

In addition to testing the accuracy of our sensors using a combined set of features, we decided to test each feature independently to see which features extracted from our sensor data were the strongest. Table 4 shows our results using a 15 second window on the sensor data obtained from all of our data sets combined. The best results were 62.29% accuracy using Multilayer Perceptron with our thumb-pressure mean feature. We experimented with different combinations of sensor features to choose the final feature-set combination for our system (described in section Sensor Features). When comparing tables 3 and 4, we can see that a multi-sensor approach helped increase our performer recognition accuracy by 12.86% (from 62.29% [thumb mean alone], to 75.15% [all sensor features]) on all data sets using Multilayer Perceptron.

	MLP (%)	SMO (%)	Naïve Bayes (%)
Mean (T)	62.29	57.86	59.15
Variance (T)	36	36	34.71
RMS (T)	37.57	34.86	36.43
SC (T)	20.57	24.57	23.43
Mean (F)	21	20.71	20.43
Variance (F)	22	18.71	23.57
RMS (F)	27	21.14	23.71
SC (F)	20.57	24.43	20.57

Table 4 - Accuracy Achieved using Individual Sensor Features on All Data Sets, T=Thumb F=Fret (15 Second Windowing)

5.4 Multimodal Results

	Exerc. (%)	*Yaman* (%)	Improv (%)	All (%)
MLP	100	100	100	100
SMO	97.33	100	92.5	86.14
Naïve Bayes	85.33	100	93.5	67

Table 5 - Accuracy Achieved using Multimodal Data (15 Second Windowing)

The results in this section were achieved by combining both the audio and sensor features into a multimodal database. Table 5 (above) shows the accuracy of the same three classifiers applied to all of the data sets as in tables 2 and 3. Multilayer Perceptron proved to be our best classifier here, yielding 100% accuracy on all data sets. While our experiments using either the audio data/features only or the sensor data/features only were satisfactory, combining them together into a multimodal database proved to be the most effective solution for performer recognition. This corroborates the use of a multimodal approach to improve systems for musical practice information acquisition.

5.5 Windowing results

Table 6 shows the accuracy of our system using Multilayer Perceptron over a variety of window periods. Our machine-learning experiments yielded the best results with a window size of 15 seconds.

The decrease in reliability of the computer's ability to perform musician recognition around our 15 second window sweet-spot can be attributed to a variety of factors. As the window size decreases, size of the training set increases accordingly, however, as a result, each feature describes a smaller (and perhaps less meaningful) piece of music. For example, the mean value derived from the thumb pressure sensor at 5 second windows, while providing more "mean values" than larger window sizes may not provide a large enough chunk of music for the extracted mean to be meaningful. Likewise, 30 second intervals may not be an appropriate representation of the actual thumb-pressure mean because the mean was not determined frequently enough. Furthermore, (with the one exception of our sensor corpus at 10-second windows), the accuracy identification at 10-seconds, 5-seconds, and 3-seconds, reduces. This suggests that the features need to be determined over a longer window period to allow enough information (samples) to be examined for an accurate representation of the feature.

Window Size (seconds)	Audio only (%)	Sensor only (%)	Multimodal (%)
60	84.57	72	93.14
30	85.71	74.57	96.28
15	85	75.15	100
10	84.09	79.24	98.85
5	82.33	76.38	97.76
3	74.97	72.43	96.29

Table 6 – Identification Accuracy of Sensors vs. Audio vs. Multimodal Fusion using a combined corpus from all data sets (at various window periods)

5.6 Training and Testing on Different Data Sets

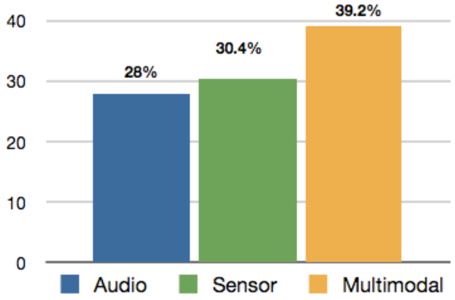


Chart 1 - Audio vs. Sensor vs. Multimodal Accuracy Achieved for Improv data set after training with Exercise and *Yaman* data sets.

For this experiment we trained the machine learning algorithms using the Exercises and *Yaman gat* data sets, then attempted player recognition on the improvisation data. This is a much more difficult, but perhaps a more "real" situation, in which we trained the system on a particularly defined set of data and asked it to classify freely improvised playing.

Chart 1 (above) is a bar graph comparing the results we achieved for audio-only, sensor-only, and our multimodal approach, using the Multilayer Perceptron classifier in each case. In contrast to previous trends, sensors features alone had a slightly higher success rate than audio features alone (30.4% accuracy vs. 28% accuracy). But as with previous experiments however, our multimodal approach was the most successful, with an accuracy rate of 39.2%. Although these results are far from perfect, they are very encouraging in that a multimodal approach improves successful musician recognition even in this more difficult case. The fact that the audio features alone performed only 8% more accurately than chance indicates that there may be room for improvement with our system's audio features.

6. Conclusion

We have developed a multimodal system to gather performance data from musicians and extract features that relate elements of their technique and personal style. Promising initial results in performer recognition show that our system is able to capture something meaningful about these performances and suggest that our system may also be useful for other tasks.

We hope that gathering a larger data set, with data from additional sensors, will allow future work to more accurately perform player identification. In addition, we hope to explore additional audio features to make our multimodal system more robust. We have built our framework on this concept, and our software suite, Super Recorder, already supports additional sensors on the ESitar. We hope to advance our work both on the ESitar, as well as to other instruments so that our research will reach and benefit a wider audience of training musicians.

Even with our simple features, the success at performer ID of our multimodal machine learning approach suggests that this approach captures something essential about an individuals sitar playing,. Much like a training athlete records workout routines, weight-lift increases, or fastest running times, we hope to use our system to quantify information about a player's performance. For example, because our data were recorded synchronously with a metronome, we can compare, for each note, when the player's left hand pressed the fret versus when the note was actually played (using an audio-based onset detector) versus when the note was expected to be played (based on tempo). Additionally, by examining the thumb sensor data synchronously recorded, it may be possible to determine if the player used the correct plucking stroke (up/down). This proposes many exciting questions regarding a musician's playing technique. How close to the expected pluck time did the player actually play the note? Was it early? Late? How does the players' performance change at different speeds (tempos) or musical contexts (practice exercises, improvisation…etc)? Did the musician use the correct stroke when playing? All of these questions are important to a musician's practice as he/she trains to become a more accurate, and expressive musician, and a tool to quantify and track these empirical aspects of musical performance could be quite valuable for musicians' development. In addition to developmental-centric information about a musicians playing, reliably extracting stylistic elements from a player's performance is also a possibility of future research using this approach. Because two similarly skilled musicians can develop individually stylistic elements to their playing, it is also possible that by examining our data, stylistic differences can be identified through empirical means. We believe our results showing the ability of the computer to perform musician recognition suggests the data sets recorded provide insight into the answers of these questions.

While this research attempts to test our multimodal approach using traditional classification techniques, we have also tried to create an umbrella in which under, a future of interactive learning can take place through a musician's interaction with intelligent machines. We have provided empirical data that shows it is possible to create unobtrusive, HCI-minded instruments that require very little adjustment on the part of the musician, while opening the doors to a world of possibilities through the use of HCI and machine learning.

7. Acknowledgements

We would like to thank Ustad Aashish Khan for his kind insight into the theory, technique, and performance of North Indian classical music. We would also like to thank Suman Laha, Rodrigo Restrepo Pabon, and William Marsh for playing sitar in our experiments. Further thanks to George Tzanetakis for training on machine learning techniques.

8. References

[1] Kapur, A., *Digitizing North Indian Music: Preservation and Extension using Multimodal Sensor Systems, Machine Learning and Robotics*. 2008: VDM Verlag.

[2] Kapur, A., Percival, G., Legrange, M. et. al., *Pedagogical Transcription For Multimodal Sitar Performance.* in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR). 2007.*

[3] Kapur, A., R.I. McWalter, and G. Tzanetakis. *New Music Interfaces for Rhythm-Based Retrieval*. *ISMIR*. 2005. London, England.

[4] Khan, A. A., Ruckert, G., "The Classical Music of North India," Munshiram Manoharlal Publishers Pvt. Ltd, New Delhi India 1998

[5] Machover, T., *Hyperinstruments - A Progress Report 1987 - 1991*. 1992, MIT Media Labratory.

[6] Ramirez, R., et al. *Performer Identification in Celtic Violin Recordings*. in *ISMIR*. 2008.

[7] Ramirez, R., et al., *Performance-Based Interpreter Identification in Saxophone Audio Recordings.* IEEE Transactions on Circuits and Systems for Video Technology 2007. **17**(3): p. 356-364.

[8] Stamatatos, E. and G. Widmer, *Automatic identification of music performers with learning ensembles.* Artif. Intell., 2005. **165**(1): p. 37-56.

[9] Stamatatos, E. *Quantifying the Differences between Music Performers: Score vs. Norm*. in *Proceedings of the International Computer Music Conference (ICMC)*. 2002.

[10] Stamatatos, E. and G. Widmer. *Music Performer Recognition Using an Ensemble of Simple Classifiers*. in *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI 02)*. 2002.

[11] Stamatatos, E. *A Computational Model for Discriminating Music Performers*. in *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*. 2001.

[12] Wang, G., *The ChucK Audio Programming Language: A Strongly-timed and On-the-fly Environ/mentality*. 2008, Princeton University.

[13] Widmer, G., *Using AI and Machine Learning to Study Expressive Music Performance: Project Survery and First Report.* AI Communications, 2001. **14**.

[14] Witten, I. H., E.F., *Data Mining: Practical machine learning tools with Java implementations*. 2 ed. 2000.

[15] Wright, M., A. Freed, and A. Momeni, *OpenSound Control: state of the art 2003*, in *Proceedings of the 2003 conference on New interfaces for musical expression*. 2003, National University of Singapore: Montreal, Quebec, Canada.