Multimodal Sensor Analysis of Sitar Performance: Where is the Beat?

Manjinder Singh Benning[1,3] Ajay Kapur[1,3] Bernie C. Till[2] George Tzanetakis[1]

University of Victoria Music Intelligence and Sound Technology Interdisciplinary Centre (MISTIC)[1]
University of Victoria Assistive Technology Team (UVATT) [2]
KarmetiK Technology (A Division of KarmetiK LLC) [3]
Victoria, British Columbia, Canada

***Abstract*— In this paper we describe a system for detecting the tempo of sitar performance using a multimodal signal processing approach. Real-time measurements are obtained from sensors on the instrument and by wearable sensors on the performer's body. Experiments comparing audio-based and sensor-based tempo tracking are described. The real-time tempo tracking method is based on extracting onsets and applying Kalman filtering. We show how late fusion of the audio and sensor tempo estimates can improve tracking. The obtained results are used to inform design parameters for a real-time system for human-robot musical performance.**

Keywords—musical multi-modal sensors systems; tempo tracking; kalman filter; musical robotics

Topic area—Modeling of Multimedia Interaction

I. INTRODUCTION

The "intelligence" of interactive multimedia systems of the future will rely on capturing data from humans using multimodal systems incorporating a variety of environmental sensors. Research on obtaining accurate perception about human action is crucial in building intelligent machine response. This paper describes experiments involving testing the accuracy of machine perception in the context of music performance. The goal of this work is to develop an effective system for human-robot music interaction.

Conducting these types of experiments in the realm of music is obviously challenging, but fascinating at the same time. This is facilitated by the fact that music is a language with traditional rules, which must be obeyed to constrain a machine's response. Therefore the evaluation of successful algorithms by scientists and engineers is feasible. More importantly, it is possible to extend the number crunching into a cultural exhibition, building a system that contains a novel form of artistic expression, which can be used on stage.

More specifically, this paper describes a multimodal sensor capturing system for traditional sitar performance. Sensors for extracting performance information are placed on the instrument. In addition wearable sensor are placed on the human performer. A robotic drummer has been built to accompany the sitar player. In this research, the authors ask the question: How does one make a robot perform in tempo with the human sitar player?

Analysis of accuracy of various methods of achieving this goal is presented. For each signal (sensors and audio) we extract onsets that are subsequently processed by Kalman filtering [1] for tempo tracking [2]. Late fusion of the tempo estimates is shown to be superior to using each signal individually. The final result is a real-time system with a robotic drummer changing tempo with the sitar performer in real-time. For the remainder of the introduction we describe representative related work.

The goal of this paper is to improve tempo tracking in human-machine interaction. Tempo is one of the most important elements of music performance and there has been extensive work in automatic tempo tracking on audio signals [3]. We extend this work by incorporating information from sensors in addition to the audio signal. Without effective real-time tempo tracking, human-machine performance has to rely on a fixed beat, making them sound dry and artificial. The area of machine musicianship is the computer music communities' term for machine perception. Robert Rowe (who also coined the term machine musicianship) describes a computer system which can analyze, perform and compose music based on traditional music theory [4]. Other systems which have influenced the community in this domain are Dannenberg's score following system [5], George Lewis's Voyager [6], and Pachet's Continuator [7]. The idea of extending traditional acoustic instruments with sensors to capture performance information has been explored in [8].

There are few systems that have closed the loop to create a real live human/robotic performance system. Audiences who experienced Mari Kimura's recital with the LEMUR GuitarBot [9] can testify to its effectiveness. Gil Weinberg's robotic drummer Haile [10] continues to grow in capabilities to interact with a live human percussionist [11]. Our system is different as it involves multimodal sensor design to obtain improved accuracy for machine perception.

Section II presents the experimental procedure administered including details about the sensor capturing systems, wearable sensors and the robotic drummer. Section III describes the results of the experiments influencing design decisions for the real-time system. Section IV contains concluding remarks, including information on how to view online videos of the robot playing in tempo with a live sitar performer.

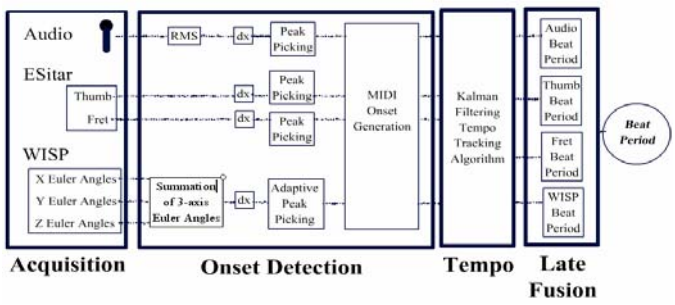

Figure 1 - Block Diagram of the system.

II. METHOD

There are four major processing stages in our system. A block diagram of the system is shown in Figure 1. In the following subsection we describe each processing stage from left to right. In the acquisition stage performance information is collected using audio capture, two sensors on the instrument and a wearable sensor on the performers body. Onsets for each signal separately are detected after some initial signal conditioning. The onsets are used as input to 4 Kalman filters used for tempo tracking. The estimated beat periods for each signal are finally fused to provide a single estimate of the tempo.

A. Acquisition and Rendering of Music Performance

This section describes the tools used to capture data from the human performer including the Electronic Sitar and wearable sensors known as the WISP, as well as the robotic drumming system known as the MahaDeviBot.

Figure 2 - Multimodal Sensors for Sitar Performance Perception.

1)The ESitar

The Electronic Sitar (ESitar) is a custom made sitar created to encase multimodal sensor technology. The sitar is the prevalent stringed instrument of North Indian classical music traditionally employed to perform ragas. The sitar is characterized by a gourd resonating chamber, sympathetic strings, and curved frets. The second author's initial work on transforming the sitar into a hyperinstrument is described in [12], which serves as a source to gain a more detailed background on traditional sitar performance technique, the sitar's evolution with technology, and initial experimentation and design in building a controller out of a sitar. The ESitar has a variety of sensors. The ones relevant to this work are a resistor network to obtain fret/neck position, and a force sensing resistor to capture pressure of the right hand thumb during performance. In a melodic stringed instrument such as the sitar rhythm is reflected by the strumming of the right hand (captured by the thumb sensor) as well as the finger changes of the left hand based on the melody played (captured by the fret sensor).

2)The WISP

The Wireless Inertial Sensor Package (WISP) [13] was designed by the third author specifically for the task of capturing human body movements. With the wireless WISP, the performer is free to move within a radius of about 50m with no other restrictions imposed by the technology such as weight or wiring. For this experiment a WISP was attached to the top of the right hand wrist to capture the movements of strumming.

The WISP is a highly integrated IMU with on-board DSP and radio communication resources. It consists of a triaxial differential capacitance accelerometer, a triaxial magnetoresistive bridge magnetometer, a pair of biaxial vibrating mass coriolis-type rate gyros, and a NTC thermistor. This permits temperature-compensated measurements of linear acceleration, orientation, and angular velocity. The first generation prototype of WISP, shown in Figure 3 next to a Canadian two-dollar coin, uses a 900 MHz transceiver with a 50Kb/s data rate. With a volume of less than 13cm^3 and a mass of less than 23g, including battery, the unit is about the size of a somewhat large wristwatch. The WISP can operate for over 17 hours on a single 3.6V rechargeable Lithium cell, which accounts for over 50% of the volume and over 75% of the mass of the unit. As can be seen in Figure 3, the small size and flat form-factor make it ideal for unobtrusive, live and on-stage, real-time motion-capture.

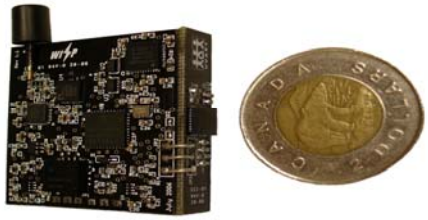

Figure 3 - Wireless Inertial Sensor Package (WISP)

3)The MahaDeviBot

We use the extracted tempo to control a robotic Indian drummer, known as *MahaDeviBot* that performs with a human sitar player in real-time. The robot has multiple arms performing a number of different instruments from India, including frame drums, shakers, bells, and cymbals. As shown in Figure 4, the lower level of the robot has four arms, each playing its own frame drum and actuated by a custom built solenoid system [14].

Figure 4 - MahaDeviBot Robotic Indian Drummer.

B. Data Collection

For our experiments we recorded a data set of a performer playing the ESitar with a WISP on the right hand. Audio files were captured at a sampling rate of 44100 Hz. Thumb pressure and fret sensor data synchronized with audio analysis windows were recorded with Marsyas[15] at a sampling rate of 44100/512 Hz using Musical Instrument Digital Interface (MIDI) streams from the ESitar. Orientation data for the Open Sound Control (OSC) [16] streams of the WISP were also recorded.

While playing, the performer listened to a constant tempo metronome through headphones. 104 trials were recorded, with each trial lasting 30 seconds. Trials were evenly split into 80, 100, 120, and 140 BPM, using the metronome connected to the headphones. The performer would begin each trial by playing a scale at a quarter note tempo, and then a second time at double the tempo. The rest of the trial was an improvised session in tempo with the metronome.

C. Onset Detection

Onset detection algorithms were applied to the audio and sensor signals to gather periodicity information of the music performance. As seen in Figure 1, the RMS energy of the audio signal and the sum of the WISP 3-axes Euler angles are calculated while the values of the thumb and fret sensor are used directly. A peak-picking algorithm is applied to the derivatives of each signal to find onset locations. An adaptive peak-picking algorithm is applied on the WISP data to compensate for the large variability in wrist movement during performance.

Each sensor captures different aspects of the underlying rhythmic structure therefore the onset streams are not identical in onset locations and phase. However we can expect that the distance between successive onsets will frequently coincide with the underlying tempo period and its multiples. In order to detect this underlying tempo period we utilize a real-time Kalman filtering algorithm for each sequence of onsets transmitted as MIDI signals.

D. Switching Kalman Filtering

Real-time tempo tracking is performed using a probabilistic Particle Filter. The algorithm tests various hypotheses of the output of a switching Kalman Filter against noisy onset measurements providing an optimal estimate of the beat period and beat [17]. Noisy onset measurements, extracted from the various sensor streams, are used as input to a real-time implementation of the tempo tracking algorithm [2]. In order to model the onset sequence we use a linear dynamical system as proposed in Cemgil [17]. The state vector x_k describing the system at a certain moment in time consists of the onset time τ_k and the period Δ_k defined as follows:

$$x_k = \begin{pmatrix} \tau_k \\ \Delta_k \end{pmatrix} = \begin{pmatrix} 1 & \gamma_k \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tau_{k-1} \\ \Delta_{k-1} \end{pmatrix} + w_k$$

As can be seen the current state vector depends on the previous state vector x_{k-1} modified by a switching variable representing different rhythmic units (γ_k) and noise model (w_k) that takes into account deviations from the ideal sequence of onsets.

Based on the above linear dynamical system the optimal sequence of tempo periods can be estimated using a Kalman filtering based approach. For more details please refer to [17].

In the following section the accuracy of the four estimated beat period streams is evaluated. In addition we show that late fusion of the streams can significantly improve tempo detection accuracy.

III. Experimental Results

TABLE I. Comparison of Aquisition Methods

Signal	Tempo (BPM)			
	80	***100***	***120***	***140***
Audio	46%	85%	86%	80%
Fret	27%	27%	57%	56%
Thumb	35%	62%	75%	65%
WISP	50%	91%	69%	53%
LATE FUSION:				
Audio/WISP/Thumb/Fret	45%	83%	89%	84%
Audio/WISP/Thumb	55%	88%	90%	82%
Audio/ WISP	58%	88%	89%	72%
Audio/Thumb	57%	88%	90%	80%
WISP/Thumb	47%	95%	78%	69%

Table I shows the percentages of frames for which the tempo was correctly estimated. Tempo estimates are generated at 86Hz resulting in approximately 2600 estimates/30 second clip in the dataset. From the percentages of Table I, we can conclude that when using a single acquisition method, the WISP obtained the best results at slower tempos, and the audio signal was best for faster tempos. Overall, the audio signal performed the best as a single input, whereas the fret data provided the least accurate information.

When looking carefully through the detected onsets from the different types of acquisition methods, we observed that they exhibit outliers and discontinuities at times. To address this problem we utilize a late fusion approach where we consider each acquisition method in turn for discontinuities. If a discontinuity is found, we consider the next acquisition method, and repeat the process until either a smooth estimate is obtained or all acquisition methods have been exhausted. When performing late fusion the acquisition methods are considered in the order listed on bottom half of Table 1.

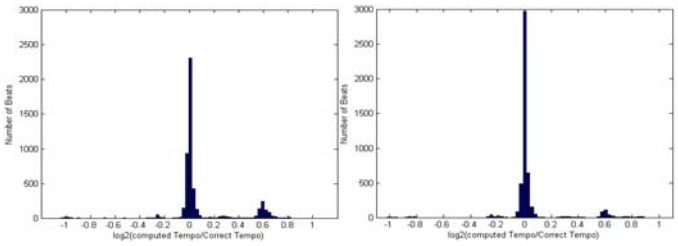

Figure 5 - Normalized Histogram of Tempo Estimation of Audio (left) and Fused Audio and Thumb (right)

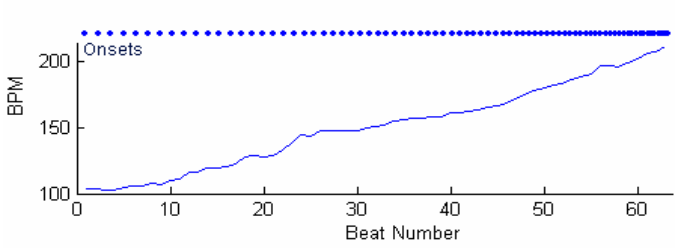

Figure 6 - Kalman Tempo Tracking with decreasing onset periods.

By fusing the acquisition methods together, we are able to get more accurate results. At 80 BPM, by fusing the information from WISP and the audio streams, the algorithm generates stronger results then either signal on its own. When all the sensors are used together, the most accurate results are achieved at 140 BPM, proving that even the fret data can improve accuracy of tempo estimation. Overall, the information from the audio fused with the thumb sensor was the strongest.

Figure 5 shows histograms of the ratio between the estimated tempo and the correct tempo. The ratios are plotted on a $\log_2$ scale where the zero point indicates correct tempo while -1, and +1 indicate half and double errors respectively. Errors of 3/2 noticed at 0.6 on the $\log_2$ scale can be attributed to the tempo tracker falsely following triple meter onsets [3]. Figure 5 shows that a greater accuracy can be achieved by fusing the audio stream with the thumb sensor stream.

IV. Conclusions and Future Work

Our results confirm that sensors are important in machine perception for musical performance. They help capture different periodicity measurements from various gestures during performance. We show that Kalman Filtering is useful for tempo tracking for both audio and sensor data. Moreover, fusing multimodal acquisition methods produces the most accurate perception results. Results from our experiments informed design parameters for our real-time system for sitar player performing with the *MahaDeviBot* in real time. It is important to note that although the experimental results presented in this paper for practical reasons deal with constant tempos one of the main advantages the Kalman filtering method is the ability to track varying tempo in real-time, as shown in Figure 6. A demonstration can be found online[1].

There is a number of directions for future work. We plan to gather more gesture data from the performer by adding more sensors to the instrument and attaching multiple WISPs in a variety of locations on the human body. We also plan to improve onset detection with more adaptive techniques. Tweaking parameters of the Kalman filter adaptively could also achieve more accurate results. Our main goal is to have a modular system that can work for any performer playing any musical instrument.

Acknowledgments

We would like to thank Tim van Kasteren for his software for Kalman filtered tempo tracking. Thanks to W. Andrew Schloss and Peter Driessen for support.

References

[1] R. G. Brown and P. Y. C. Hwang, *Introduction of Random Signals and Applied Kalman Filtering.*: John Wiley & Sons, Inc. , 1992.

[2] T. v. Kasteren, "Realtime Tempo Tracking using Kalman Filtering," in *Computer Science*. vol. Masters Amsterdam: University of Amsterdam, 2006.

[3] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An Experimental Comparison of Audio Tempo Induction Algorithms," *IEEE Transactions on Speech and Audio Processing* vol. 14, 2006.

[4] R. Rowe, *Machine Musicianship*. Cambridge, MA: MIT Press, 2004.

[5] R. B. Dannenberg, "An On-line Algorithm for Real-Time Accompaniment," in *International Computer Music Conference (ICMC)*, Paris, France, 1984, pp. 193-198.

[6] G. Lewis, "Too Many Notes: Computers, Complexity and Culture in Voyager," *Leonardo Music Journal*, vol. 10, pp. 33-39, 2000.

[7] F. Pachet, "The Continuator: Musical interaction with Style " in *ICMC*, Goteborg, Sweden, 2002.

[8] T. Machover and J. Chung, "Hyperinstruments: Musically Intelligent and Interactive Performance and Creativity Systems," in *ICMC* 1989, pp. 186-190.

[9] E. Singer, K. Larke, and D. Bianciardi, "LEMUR GuitarBot: MIDI Robotic String Instrument," in *International Conference on New Interfaces for Musical Expression (NIME)*, Montreal, Canada, 2003.

[10] G. Weinberg, S. Driscoll, and M. Parry, "Haile - A Preceptual Robotic Percussionist," in *ICMC*, Barcelona, Spain, 2005.

[11] G. Weinberg, S. Driscoll, and T. Thatcher, "Jam'aa - A Middle Eastern Percussion Ensemble for Human and Robotic Players," in *ICMC*, New Orleans, 2006, pp. 464-467.

[12] A. Kapur, A. Lazier, P. Davidson, R. S. Wilson, and P. R. Cook, "The Electronic Sitar Controller," in *NIME*, Hamamatsu, Japan, 2004.

[13] B. C. Till, M. S. Benning, and N. Livingston, "Wireless Inertial Sensor Package (WISP)," in *NIME*, New York City, 2007.

[14] E. Singer, J. Feddersen, C. Redmon, and B. Bowen, "LEMUR's Musical Robots," in *NIME*, Hamamatsu, Japan, 2004.

[15] G. Tzanetakis and P. R. Cook, "Marsyas: a Framework for Audio Analysis," *Organized Sound*, vol. 4, 2000.

[16] M. Wright, A. Freed, and A. Momeni, "OpenSound Control: State of the Art 2003," in *NIME*, Montreal, Canada, 2003.

[17] T. Cemgil, "Bayesian Music Transcription." vol. Ph.D. Netherlands: Radboud University of Nijmegen, 2004.

[1] http://www.karmetik.com (Technology → Robotics)