PROTOTYPE FOR MOTION INITIATED MUSIC ENSEMBLE WITH SENSORS (MIMES)

Dania Murad [1] **Ye Fenyi**[3] **Arun Shenoy**[1]
Michael Barone[1] **Simon Lui**[2] **Ye Wang**[1]
[1] National University of Singapore, Singapore
[2] Singapore University of Technology and Design, Singapore ; [3] Fudan University, China
daniamurad@comp.nus.edu.sg, wangye@comp.nus.edu.sg

ABSTRACT

We introduce *MIMES* (Motion Initiated Music Ensemble with Sensors); a computational audio system which uses hand gestures for music generation. By using inertial sensors from smart watches, the system is able to trigger instrumental sounds with simple hand gestures. The simplicity of gestures enables novices to learn fundamental aspects of music generation with a certain degree of effectiveness in terms of harmony and rhythm, before they master motor skills required for conventional instruments and group performance. The ability of *MIMES* to support several performers as a group is presented as a key feature of the proposed system. We describe signal processing and machine learning techniques required to process inertial sensors inputs to detect different gestures. Gesture recognition accuracy is evaluated, and future work describes how additional gestures will be incorporated to drive a music composing system.

1. MOTIVATION

Technological advances in inertial sensors have made gesture recognition systems more feasible to implement than ever before, providing high resolution signals for estimating the dynamics of motion in real-time. As a result, gesture recognition applications are expanding in use and scope including entertainment, education, and health care. The motivation for developing the system is to aid performers who have yet to develop a sense of rhythm or theoretical background to be able to participate in a musical group performance. Group performances have indeed existed for millenia, but performed by artists who have dedicated years to the mastery of specific instruments. Here we are targeting non-musicians to be able to achieve the same proficiency to some extent without the requirement of fine motor control to generate clear sounds of a specific instrument. Given that music is inherently a social experience, for musicians and listeners alike, a system designed to teach beginners that can incorporate gestures from multiple performers simultaneously is potentially useful for music education and creation. The system can also be used for Therapeutic Instrumental Music Performance (TIMP) which involves using instrument playing for rehabilitation of motor skills. To that end, we have developed a real-time framework for music composition, based on an architecture which uses hand gestures to dynamically render musical arrangements from multiple performers.

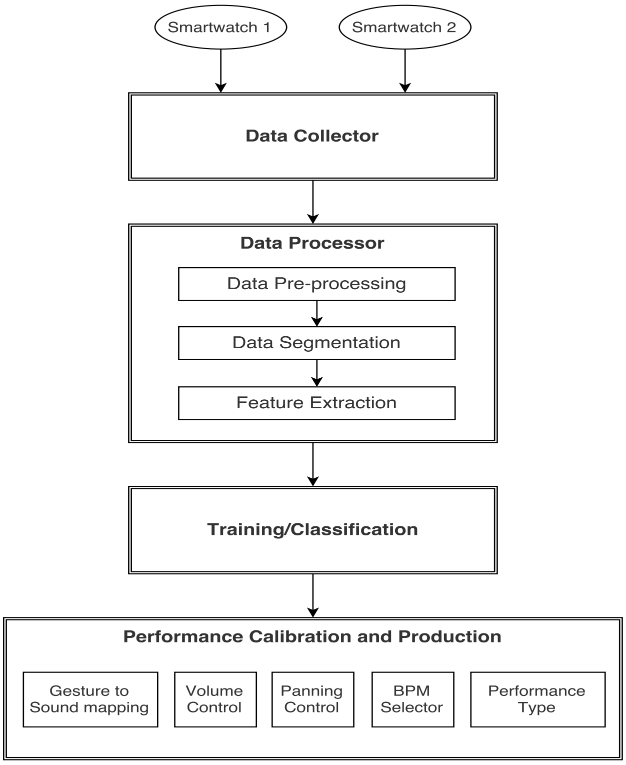

Figure 1. MIMES Architecture Workflow

2. METHODOLOGY

Figure 1 describes the *MIMES* processing workflow: i) *Data Collector*; ii) *Data Processor*; iii) *Training and Classification* iv) *Performance Calibration and Production*. The details are provided as below:

Attribution: Dania Murad, Ye Fenyi, Arun Shenoy, Michael Barone, Simon Lui, Ye Wang. “Prototype for Motion Initiated Music Ensemble with Sensors (MIMES)”, Extended abstracts for the Late-Breaking Demo Session of the 18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017.

2.1 Data Collector

The Data Collector samples accelerometer data from smart watches at 100 Hz, and transmits it to the Data Processor on a local server via Bluetooth 4.0.

2.2 Data Processor

Accelerometer data passes through many sub-processes for real-time detection of gestures. To eliminate disturbances caused by unintended movements, an averaging filter is used to remove high frequency noise from the signal. To segment a meaningful gesture from filtered signal, Euclidean distance is measured between successive accelerometer samples. Time domain features including mean, standard deviation, variance, minimum value and maximum value are calculated across each of the three axes. To extract features using dynamic time warping (DTW), a reference signal for each gesture is recorded and saved. Warping path is calculated between the reference and the newly performed gesture across all the three axes, providing $3N$ features, where N is the number of gestures defined by the performer. We defined nine gestures for evaluation, resulting in 42-dimensional feature vector.

2.3 Training and Classification

In the training phase, user defined gestures are collected and passed through pre-processing and segmentation techniques. Salient features, as described in Section 2.2, are extracted from the segmented gesture and provided as inputs to the feed-forward neural network, consisting of three layers. The number of nodes in the output layer corresponds to the desired number of user-defined gestures. For our demonstration, the model is trained with a dataset of 1080 segmented gestures (120 recordings per gesture) collected from a single performer, with nine gestures (Table 1) as labels. The new gestures performed by the users are then classified by the trained model to produce sounds for music composition.

3. PERFORMANCE CALIBRATION AND PRODUCTION

Once the network has been trained, performers can calibrate their composition by mapping any sound to the output. Performers can also select one of the two modes: non-synchronized and synchronized. In non-synchronized mode, audio playback occurs as soon as a gesture is performed; this mode is useful for performers who wants to understand the fundamentals of music in a solo performance setting, or for group of performers with a strong sense of rhythm. In synchronized mode, the system automatically corrects the minor errors in timing and enable the performers to create a musical piece that adheres to a specific tempo and rhythm. A tactile metronome is provided to them by making the smart watch vibrate with a predetermined beat. This cues performers when they should perform a gesture, minimizing asynchrony between them. For a natural sounding group performance, a wide range of musical instruments should come together with a keen sense of rhythm, harmony and melody. For this purpose, level and spatial arrangement of each element in the stereo field is also incorporated into the framework.

4. EVALUATION

Gesture classification accuracy is determined through post-hoc analysis. To test the usability of the system with the data being processed in real-time, each of the nine gestures is executed 100 times. A trial is considered successful if musical output corresponds to the appropriate gesture with above 90% confidence. Table 1 illustrates the percentage accuracy for each output class. The average accuracy across all classes is 95.9% ($SD = 3.7\%$), achieving a minimum of 89%. The music produced by the performed gestures then renders the mixed musical audio signal stream to the output device in real-time.

Gesture	Success (>90% Confidence)
Flick Up	94%
Flick Down	95%
Flick Right	91%
Flick Left	98%
Forward	89%
Backward	100%
Swipe Down	99%
Clockwise	98%
Counter-clockwise	99%

Table 1. MIMES Gesture Recognition Accuracy

5. FUTURE WORK

New gesture types will be incorporated to improve potential use cases for *MIMES*. Presently, gestures have discrete onsets and offsets; continuous gestures will be more appropriate to map to instruments that sustain tones for longer periods of time, i.e. violin. Furthermore, training the network to recognize finger gestures would significantly increase the amount of possible degrees of freedom. Lastly, because *MIMES* can generate musical sounds without fine-motor control, we foresee potential health applications for individuals with motor disabilities [1] such as Huntington's, Alzheimer's, cerebral palsy, and stroke patients.

6. REFERENCES

[1] Yao-Jen Chang, Shu-Fang Chen, and Jun-Da Huang. A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in developmental disabilities*, 32(6):2566–2570, 2011.