Gus XIA RESEARCH STATEMENT

As both a computer scientist and a musician, I design intelligent systems to *understand* and *extend* human musical expression. To *understand* means to model the musical expression conveyed through acoustic, gestural, and emotional signals. To *extend* means to use this understanding to create expressive, interactive, and autonomous agents, serving both amateur and professional musicians. Such systems provide intimate human-computer interaction and further blur the boundary between humans and machines.

Intelligent systems have already reached the level of human capabilities for many tasks, including chess and board games, driving aircraft and vehicles, and even speech and face recognition. However, these tasks are very functional; for the tasks requiring an understanding of inner human expression, computers are still far behind. I believe that the study of musical intelligence, being one of the most profound aspects of humanity, will be the next frontier of artificial intelligence.

To understand and extend musical expression, which is implicit and abstract, we need to create *artificial musicianship*. My work combines music domain knowledge with computational techniques, especially machine learning algorithms, for machine music understanding. Furthermore, I apply tools of human-computer interaction and robotics to extend human musical expression. The systems I create offer new ways for computers to participate in live performance. Such systems include: 1) an interactive artificial performer built on the understanding of music timing and dynamics, 2) an autonomous dancing robot built on the understanding of beat and music emotion, and 3) a smart music display as a bi-directional interface built on the understanding of music notation.

Interactive Artificial Performers

I create interactive artificial performers [1,2] that are able to perform expressively in concert with humans by learning musicianship from rehearsal experience. This study unifies machine learning and knowledge representation of music structure and performance skills in an HCI framework. The solution is inspired from my own rehearsal experience, where musicians become familiar with one another's performance styles in order to achieve better prediction and interaction on stage.

In particular, I consider pitch, timing, and dynamics features of musical expression and model these features across different performers as co-evolving time series that vary jointly with *hidden* (mental) states. Then, I apply *spectral learning*, a state-of-the-art machine learning technique, to learn how the time series evolve over the course of a performance. The spectral method first computes empirical moments of the high-dimensional features, then applies *singular value decomposition* on these moments to learn a compact representation of the hidden states, and finally recovers the model parameters using their relationship with the hidden states. For such a complex time-series learning task, the spectral method provides two benefits as compared to maximum likelihood estimation: it is computationally efficient and free of local optima.

Based on the trained model, the artificial performer generates an expressive rendition of a given score by interacting with human musicians. Compared with the baseline, my approach improves the timing prediction by 50 milliseconds and the dynamics (loudness) prediction by 8 MIDI velocity units on average, trained on only 4 to 8 rehearsals of the same piece of music (Figure 1).

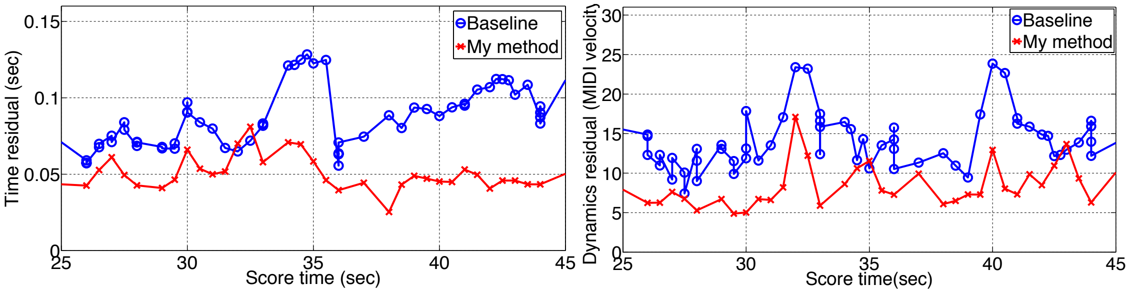

Figure 1. A comparison of the cross-validation results between the baseline and my method trained on only 4 rehearsals (smaller number is better).

This is a *very significant* improvement, as listeners can easily perceive asynchronous notes that differ by 30 milliseconds and dynamics differences of 4 MIDI velocity units. The baseline we compared with is a rule-based approach, which assumes that local tempo and dynamics are steady and the synchronization is perfect. This baseline model has been used in *automatic accompaniment* systems for over 30 years! Now, my work has demonstrated that this model is inadequate: local deviations in timing and dynamics, also known as *musical nuance*, play an important role in expressive musical interaction. Remarkably, seemingly irregular musical nuance is related to predictable hidden states, and computers are able to learn this relationship from a small number of rehearsals. An audio demo is available at *http://www.cs.cmu.edu/~gxia/demo1.pptx.*

Even when learning from the rehearsals of *different* pieces in similar music styles, my approach can still outperform the baseline. This indicates that expressive musical interaction, which involves musical nuance, follows universal patterns. In addition, computers are capable of generalizing what they have learned and applying it in new situations.

Musical expression extends beyond sound. Recently, I collaborated with world-leading humanoid music robots and extended artificial musical expression to incorporate facial expression and body gestures (Figure 2). The macro-level body gestures are designed to reflect phrase structures, while the micro-level facial expressions are designed to reflect musical nuance. A video demo is available at *https://youtu.be/AAC7wI64aBM*. As far as we know, this is the first collaborative performance between a human and a humanoid music robot with facial and gestural expression.

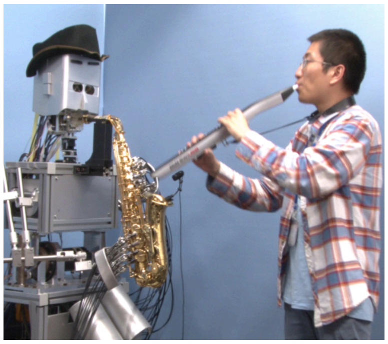

Figure 2. Human-robot music interaction between the saxophone robot and the flutist (myself).

In summary, I have created interactive artificial performers, which are able to learn musicianship from rehearsals and extend human musical expression by playing collaboratively and reacting to musical nuance with facial and body gestures. This is a big step towards artificial musical interaction at a professional level.

Autonomous Dancing Robots

Dance is another way to convey and extend musical expression, which (once again) involves issues in sensing, planning, executing, and representation. While most robot dances are still designed for a particular piece of music, I developed an autonomous dancing robot driven by music [3].

A successful robot dance performance should achieve three goals: first, the choreography should be synchronized to the beat and further reflect music emotions; second, the dance should be smooth, interesting, and non-deterministic; and third, the choreography should be safe to execute. To meet these criteria, the autonomous dancing robot is designed with three major parts: *listen, plan,* and *maintain safety*. The *listen* part uses an autocorrelation function to extract the beats and uses *support vector regression* to extract music emotion by projecting the high-dimensional acoustic features onto a 2D activation-valence emotion space. The *plan* part is the most important. It takes the beats and emotion as input guides and adopts a Markov model to generate smooth dance movements stochastically. The *maintain safety* part examines the balance and speed of the planned motions and prunes any dangerous movements. A video demo is available at https://youtu.be/mlsVL-OtBaE.

Smart Music Displays

Intelligent systems can also enhance musical expression by removing traditional obstacles in musical interaction. An example where such a system is useful relates to traditional sheet music notation, which requires too much attention from musicians. For example, musicians have to hurriedly turn the page while performing an intensive part. Even when their parts are silent, they sometimes have to count the beat throughout an entire session (musicians hate counting).

To free musicians from this tedious labor, I designed a smart music display [4,5] which automatically keeps track of the most current score location and turns pages during a performance (Figure 3). The system has three major parts: *read, listen,* and *display*. The *read* part maps a conventional score (with repeats and other structures) to a sequence of notes; the *listen* part computes the local tempo and predicts the timing of the next note and page turn; and the *display* part renders the sequence of notes on a screen in real time, using a cursor to indicate the current score location.

Moreover, the smart music display is a bi-directional user interface. Users can click on a score location and access recordings or other media files that are automatically aligned with the score [6]. This function helps musicians quickly connect notation with actual performance.

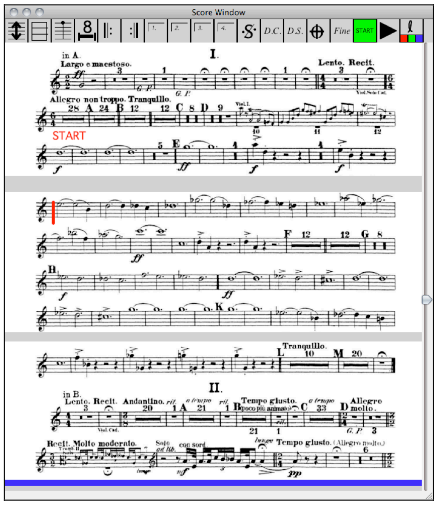

Figure 3. The smart music display. “START” shows the starting point for the playback; the vertical red cursor shows the current location.

Future Research Directions

I aim to empower intelligent systems with more profound artificial musicianship through wider interdisciplinary efforts. Such systems will be able to serve people not only through music performance but also through music education and music therapy. This section outlines three future directions; each requires a progressively deeper understanding of music and provides a larger potential impact on people’s daily lives.

Automated improvisation built on the understanding of composition

Current systems can understand music notation, beat, emotion, and musical nuance. A natural extension is to understand *music composition* in order to improvise (compose while performing) a collaborative part on top of a human performance in real time. To achieve this, we need more sophisticated machine learning techniques to address pitch, rhythm, and harmony, simultaneously. I have partly achieved this in one of my ongoing projects in collaboration with our school of music, in which I adopted a neural network to generate a piano part *offline*. A video demo is available at https://youtu.be/1GWHuzRLcbc.

With the ability to improvise and add musical nuance and gestures expressively, a robot musician can serve as a personal music partner. We can further generalize this idea and imagine a personal robot orchestra, with which even amateur musicians can hold solo concerts easily. *This is one future direction for music performance*.

Self-trained artificial performers built on a taste in music

One step further toward a profound music understanding is *taste*, the core part of human musicianship. An intelligent system with a taste in music will be able to train itself by learning from examples selectively, rather than through a passive and fully-supervised procedure. By using semi-supervised learning, especially active learning, a system can learn basic musicianship from a small number of labeled human performances and then improve itself automatically through its own experience or a vast amount of unlabeled data collected by *music information retrieval* techniques [7].

Self-trained systems will require minimal programming and calibration from humans. They will be able to use feedback to adapt their collaborative performances to different human musicians, instrument qualities, and acoustic environments. A special application is to just "copy" a performance from a concert hall and later adaptively "paste" it to our home. In other words, we can listen to "live" symphonies at home using a robot orchestra. *This is one future direction for music appreciation*.

Music education and therapy built on intelligent supervision of musicianship

Today, humans design algorithms to develop artificial musicianship; tomorrow, machines can help teach music to humans. An artificial music teacher can give us feedback at any time, making music training a lot easier and potentially much cheaper. In addition, human-robot musical interaction enables us to explore how different teaching strategies affect the way students learn, since the robot behaviors can be easily configured by a set of parameters. Each step in music training has three components: 1) an accurate judgment of the current level of musicianship, 2) a reachable next-level target, and 3) a tailored plan to reach that target. Though solving all three problems autonomously is a longer-term goal, we can combine machine learning with *human computation* (crowdsourcing) to design a music education curriculum jointly with machines.

My vision for unifying music education and therapy is inspired by *Eurhythmics,* a traditional music training method that focuses on the intrinsic relationship between body movements and musical expression. For example, a Eurhythmics instructor plays a tricky music segment on the piano; students are asked to step on the downbeats while clapping on the upbeats to show their mastery of a certain rhythm. This procedure (of training rhythmic feeling) can be turned easily into a human-computer interaction, where intelligent systems can play the piano part while evaluating the movements of the students. Moreover, this method can be adapted to physical therapy. Compared with current approaches in physical therapy for Parkinson's disease where doctors still use metronomes to help patients recover their ability to walk smoothly, an interactive process involving musical expression will be a *huge* improvement.

In summary, I design intelligent systems that "think" about music like humans — to utilize past experiences and to interact with unexpected and changing environments. I look forward to carrying on my research on music intelligence and sharing ideas on further evolving musical expression with brilliant colleagues. Let's envision a more expressive, interactive, and musical world, and make it so.

References:

[1] **G. Xia**, R. Dannenberg "Spectral Learning for Expressive Interactive Ensemble Music Performance", in *Proc. 16th International Conference on Music Information Retrieval*, Malaga, October 2015.

[2] **G. Xia**, R. Dannenberg "Duet Interaction: Learning Musicianship for Automatic Accompaniment", in *Proc. 15th The International Conference on New Interfaces for Musical Expression*, Baton Rouge, June 2015.

[3] **G. Xia**, J. Tay, R. Dannenberg, M. Veloso "Autonomous Robot Dancing Driven by Beats and Emotions of Music", *in Proc. 12th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Valencia, June 2012.

[4] D. Liang, **G. Xia**, R. Dannenberg "A Framework for Coordination and Synchronization of Media", in *Proc. 11th The International Conference on New Interfaces for Musical Expression*, Oslo, May 2011, pp. 167-172

[5] R. Dannenberg, N. Gold, D. Liang, **G. Xia** "Active Scores: Representation and Synchronization in Human-Computer Performance of Popular Music", in *Computer Music Journal*, 38 (2) (Summer2014). 2013

[6] **G. Xia**, D. Liang, R. Dannenberg, M. Harvilla "Segmentation, Clustering, and Display in a Personal Audio Database for Musicians", in *Proc. 12th International Conference on Music Information Retrieval*, Miami, October 2011, pp.139-144.

[7] **G. Xia**, T. Huang, M. Yifei, R. Dannenberg, C. Faloutsos "MidiFind: Similarity Search and Popularity Mining in Large MIDI Databases", in *Music Sound and Motion*, 2014, pp. 259 -276.