Chronicles of a Robotic Musical Companion

Mason Bretan
Georgia Institute of Technology
840 McMillan St.
Atlanta, GA
pbretan3@gatech.edu

Gil Weinberg
Georgia Institute of Technology
840 McMillan St.
Atlanta, GA
gilw@gatech.edu

ABSTRACT

As robots become more pervasive in the world we think about how this might influence the way in which people experience music. We introduce the concept of a "robotic musical companion" (RMC) in the form of Shimi, a smartphone enabled five degree-of-freedom (DoF) robotic platform. We discuss experiences individuals tend to have with music as consumers and performers and explore how these experiences can be modified, aided, or improved by the inherent synergies between a human and robot. An overview of several applications developed for Shimi is provided. These applications place Shimi in various roles and enable human-robotic interactions (HRIs) that are highlighted by more personable social communications using natural language and other forms of communication.

Keywords

musical robotics, human robotic interaction

1. INTRODUCTION

The use of robotics in music has produced new types of performance, novel interactive experiences, and state of the art algorithms for machine listening and generative purposes. By exploring the characteristics and adhering to the constraints inherent to musical robots, researchers are able to expand upon and push the envelope of the musical experience for performers and audience members alike.

Though experiencing music often consists of applications which place an individual outside of the role of performer or audience member, the focus of robotics in music has centered almost entirely on the performative aspects of music. This includes functionalities such as improvisation, automatic accompaniment, and call and response. Here, we introduce the idea of a "robotic musical companion" (RMC) and explore the notion of using robots to enhance or modify several instances of the musical experience encompassing performance, recommendation, general listening, composition, and education. We examine the possibilities pertaining to music consumption, perception, and expression through a robotic interface by describing several applications developed for Shimi, our smart-phone enabled creature-like robotic platform. These applications address some of these unexplored musical experiences in the context of human-robotic interaction (HRI) and are underscored by more personable social communications.

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

2. WHAT IS A ROBOTIC MUSICAL COMPANION?

2.1 Robots in Music

Robotic musicians are most commonly presented as a set of individual actuators often playing percussive instruments [10, 11]. The physical designs enable activation of natural acoustic resonators and provide useful visual cues. Some systems include additional mechanics to create completely embodied systems which take on humanoid [13] or creature-like forms [9].

Several of the designs for such systems are motivated by aesthetics and sound quality. Others are developed with specific characteristics that offer practical functions and utilities not necessarily related to auditory perceptions, yet are still important to music. The robotic marimba player, Shimon, explored the social interactions in music performance by incorporating a moving head and functional eye to respond to both auditory and physical (such as motion and gestures) components of a human's performance. Shimon achieves this through physical motions such as beat synchronized periodic gestures, nodding, and manipulating its gaze. The use of gestures in such a manner not only provides increased enjoyment and entertainment value for an audience, but is also useful for conveying the system's understanding of the music and underlying computational processes. This allows audience members and interacting musicians to better interpret the robot's intentions [3].

2.2 Defining a Robotic Musical Companion

The robotic systems previously mentioned are instrumentalists and their interactive experiences are constructed to address the performative aspects of music. While these types of musical interactions are certainly important and should be supported by an RMC, other scenarios in which people commonly interact with or experience music should also be considered. For example, music is frequently consumed in the form of general listening through the radio or desktop and mobile applications such as iTunes, Pandora, and Spotify [1]. Several systems also exist to recommend such as last.FM and tasteKid [2]. These applications utilize methods involving automated artificial intelligences or provide interfaces encouraging users to suggest music to friends.

Other musical experiences related to playing instruments outside of the context of performance includes instrumental practice and composition. Musicians often practice with metronomes or in front of peers and educators to receive

[1] http://www.apple.com/itunes/, http://www.pandora.com/, https://www.spotify.com/
[2] http://www.last.fm/, http://www.tastekid.com/

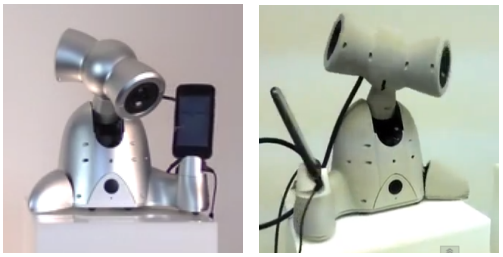

Figure 1: Shimi is a robotic musical companion that employs the "dumb robot, smart phone" approach and engages users with a variety of musical applications.

feedback. Composers often play instruments during the writing process. Additionally, compositional aids in the forms of computer interfaces are becoming more pervasive and even allow novices to develop creative and inspiring arrangements and compositions [5].

Many of these applications are easy to use and their interfaces have been optimized for the best user experience. We cannot guarantee that implementing these applications through an HRI will produce a better or worse user experience. Rather, we hope the new types of interaction made available by robotics will pique human interest, inspire human creativity, and facilitate tasks that would otherwise be dull, tedious, or difficult through a non-sociable computer interface.

Though a primary objective in developing an RMC is to explore the existing and traditional musical experiences in the context of an HRI, however, the use of a robotic interface invites new scenarios for which people can consume music. Music can be leveraged to serve additional purposes for what have been typically non-musical tasks in robotics. For example, a common research area in the field of HRI includes methods for robots to clearly communicate underlying computational and mechanical processes, system performance, and even higher level affective states. Sonifying these pieces of information with music can lead to the development of more expressive and engaging robots. Later we will describe how Shimi uses music to exhibit and communicate emotion, empathy, and comprehension.

We define an RMC as a physical agent which encourages social and personable communications to achieve fun and entertaining musical interactions, while, also utilizing music as a tool for typically non-musical tasks. In essence the robot should generate, consume, and demonstrate a knowledge of music. If the robot has a message make it sing; if it finds a beat make it dance.

3. DEVELOPING SHIMI, A ROBOTIC MUSICAL COMPANION

3.1 Design

3.1.1 Physical and Functional Design

Shimi's design is based off of the "dumb robot, smart phone" (DRSP) approach in which a mobile phone is appropriated to perform the computation, sensing, and high-level motion control for a simple robotic platform [8]. Unlike many of the existing robotic musicians which may contain a series of powerful motors, solenoids, sensors, and computers, Shimi's design is simple. It has only five degrees of freedom (DoFs) that are used to move its head, foot, and hand (which serves as the docking station for the mobile phone) and includes three speakers (one woofer, two tweeters).

The DRSP approach provides users with a familiar and easy to use interface. However, in sacrificing more complicated mechanics for a smaller portable structure, the ability for Shimi to play acoustic instruments is also lost. Two major advantages of the use of robotics in music have been the production of natural acoustic sound and the visual cues provided when playing their instruments. However, despite Shimi's inability to generate acoustic sound, the benefits of visual cues can still be yielded by cleverly synchronizing movements with the sound it is able to generate. In other words, an artificial interdependence between motion and sound can be constructed.

Such a paradigm requires developers to assume the role of composer as well as choreographer. There no longer exists an inherent or natural connection between the two elements. Instead, a correlation is explicitly designed in order to provide the necessary connection between the generated music and Shimi's physical movements. Though this requires more thought and work for the developer, it provides an opportunity for additional creativity and may even be considered advantageous. Increased emphasis can be placed on the aesthetics of the motion because the robot's movements are not constrained to what is necessary for playing an instrument and activating sound acoustically.

3.1.2 Creating Natural and Engaging Social Interactions

The design of the interactive scenarios are influenced by what Shimi is able to hear and see (through use of the smart phone's microphone and camera). Despite not having eyes, Shimi is able to move its head to effectively demonstrate gaze. We also take advantage of Google's speech recognition service and use speech as a primary form of communication making natural language processing (NLP) an important aspect of the underlying intelligence.

3.2 Functionalities and Interactive Scenarios

The following sections describe functionalities developed in our research with Shimi thus far. Many of these applications are works in progress that will continue to be expanded upon, improved, or modified. They describe a wide breadth of work and provide an overview of the types of interactive scenarios we believe can be useful for RMCs.

3.2.1 Shimi as an Interactive Speaker Dock

The built in speakers and smart phone docking station make Shimi ideal for serving as a speaker dock. However, rather than using a graphical user interface (GUI) as is common with typical mobile applications, the user can interact with a physically expressive entity. In this mode Shimi accesses the musical library on the device or uses alternative sources for retrieving music such as Spotify. The Echo Nest [3] (TEN) provides information about the music which is used for choosing songs based off of the user's input.

NLP is incorporated to support general music query and recommendation applications. In this 'query by natural language' (QbNL) mode, Shimi responds to queries such as "do you have [a specific artist or song]?" or "can you play something [with a specific genre or mood]?" Perceptron models are trained with frequently used query phrases and word spotting is used to map the user's requested parameters to features made available by TEN. This includes specific artists, song titles, genres, moods, energy, and danceability. Essentially, Shimi supports a natural language interface for TEN's query methods.

Another method of interaction includes 'query by tapping' (QbT). This music retrieval system is similar to 'query

[3]http://echonest.com/

by humming' (QbH) in which songs are chosen based off of their distinct melodies and how they compare to user input humming or singing [7]. In QbT users tap or clap a rhythmic motif and Shimi finds a song from its library which is best represented by that specific motif. Currently, our system can only function accurately on a much smaller scale (roughly 20-30 songs) than what has been successful with QbH. However, we see great potential for this to be effective in HRI. Interfaces such as mobile phones or keyboards require users to use their hands. A robotic interface such as Shimi frees users of the necessity of holding a device or pressing buttons, thus, permitting one's hands to be used for other purposes. In this case, users are free to clap as a method of communication to Shimi.

QbT works by matching rhythmic patterns to manually tagged rhythms which are considered representative of a piece. This is done using a kNN classifier based on a dynamic time warped distance metric. It is not feasible to expand QbT to millions of songs with this approach and it is not necessarily the case that a song can be successfully characterized by a single rhythmic motif. However, we believe the idea is worth pursuing as such functionality has proven popular through empirical observations during showcases and demos.

Identifying and playing the appropriate songs for these types of queries is essential in order for Shimi to exhibit a higher level musical intelligence. However, choosing the song is only the first step. Demonstrating an understanding of other low level qualities describing the music is addressed once the song begins to play.

Shimi's five DoFs provide the mobility necessary to enable dancing. Genre, beat, and section locations are features extracted from TEN that provide informational and musical cues describing how to dance. Shimi's generative motion functions enable beat synchronized movements. Shimi has a library of expressive dance moves which were specifically designed for different styles of music and can be performed at any tempo (though if a beat's duration is too short the move may not be be fully completed). As a section change occurs (i.e. verse to chorus) Shimi switches dance moves as well.

3.2.2 Shimi as a Listener

It is important for an RMC to exhibit intelligence regarding the music it is generating. It is also important for an RMC to similarly exhibit intelligence and responsiveness to music with which it is being presented. Additional capabilities were implemented to achieve this by enabling Shimi to listen and respond to a live performer. From a technical standpoint this music analysis task proves difficult because analysis must occur in real-time using only the resources available to the phone. It is also difficult to distinguish important musical content from the noise generated from Shimi's own motors. Applying a low pass filter helps to reduce some of the high frequency content Shimi produces as it moves.

Determining which characteristics in the music Shimi should listen for and respond to provides another question. As a speaker dock we found that when responding to the genre, beat location, and section location features of a song Shimi is able to demonstrate a knowledge of both higher and lower level musical parameters through dancing. In listening mode, however, we assume the live performer will take on a more improvisatory role and features describing the more immediate and subtle nuances of a performance will be necessary. We chose features to provide rough estimates of note density, pitch, loudness, and periodicity (i.e. is there a strong presence of a beat).

Analysis of the audio occurs over a three second window. Note density and fundament pitch contours are computed. Though individual pitches can be useful for calculating several high level features, currently we are measuring how pitch changes over time enabling Shimi to respond to rising and falling melodic lines. Therefore we take the first order difference of the detected fundamentals and calculate the average and standard deviation for each three second window. Average and standard deviations of the absolute magnitude spectrum is used to describe the loudness of the performance.

Auto- and cross-correlation methods described by [6] are used for detecting the presence of periodicities in the performance (i.e. does the music have a beat). During solo improvisation performers often play more freely without adhering to a strict tempo. Empirically we have found that it is very noticeable when Shimi moves periodically to a beat which the listener does not also hear. To limit these false positives Shimi only aligns with a tempo if the measured beat confidence exceeds a certain threshold. Otherwise, Shimi's movements are influenced more by the other features (note density, pitch, and loudness).

3.2.3 Shimi as a Practice and Compositional Aid

Visual cues help to increase performance of synchronization tasks in music. The manner in which something moves can be manipulated to provide additional benefits. For example, visual metronomes which exhibit acceleration and deceleration can be more helpful than those which move at a constant velocity [1]. The controller for Shimi's movements was designed to support such oscillating motions and other velocity contours. This allows better perception of Shimi's movement-beat synchronization while dancing to a song and also means Shimi can be used as a visual metronome.

A drum sequencer was built for Shimi with a natural language interface for the user. The user does not place the individual notes in the sequencer, but rather verbally asks Shimi to play drum patterns of particular styles such as, "Shimi, can you play a latin beat?" The user can then tap or clap the desired tempo. Shimi plays back the drum beat at the specified tempo and synchronizes its dance gestures. This provides the user the benefits of both auditory and visual metronomic cues.

Using natural language as an interface has both advantages and disadvantages. It is an intuitive method of communication, on the contrary, it may not provide the desired low level control and might be frustrating if a request is interpreted incorrectly. Some functions are relatively simple to interpret through NLP such as those regarding tempo ("play it [in half time, in double time, 10 bpm slower, etc]"). However, other requests can have a degree of subjectivity which can be difficult to quantify, for example, "play something a little more funky."

We built in default behavioral responses for many of these more subjective requests, but there remains a significant likelihood that the default response will not meet the user's wishes. To address this we use a stochastic sequencer allowing the user to request for increased or decreased probability that a particular drum is used at specific time steps. The sequencer starts with predefined probabilities based off a specific style template (i.e. rock, swing, bossa-nova, etc.). Then a request can be made to modify the template in some way. For example, a request to "use more bass drum and less hi-hat" will result in a increased probability of a bass drum being hit and decreased probability of a hi-hat being hit. Because the sequencer is probabilistic Shimi's patterns will change over time. Shimi remembers the most recent 64 drum strikes so if the user hears something he or she is

particularly fond of, then a request can made such as, "I like that pattern, repeat it."

3.2.4 Shimi as an Emotionally Intelligent Musical Linguist

Machines which demonstrate a sensitivity to people's emotional states by responding with understandable and relatable behaviors enable more natural, engaging, and comfortable human-robotic interactions [2, 4]. The benefits of emotionally intelligent robots have been exhibited in many interactive social situations. This is because information such as compassion, awareness, and competency can be communicated through a non-conscious "affective channel" [12], resulting in zero increased cognitive load for the user. Here, we describe a functionality which enables Shimi to detect sentiments and topics in language and generate an appropriate response using music and physical gestures.

Emotion has been synthesized by machines through different facial expressions and physical movements. Though emotionally expressive physical behaviors can be synthesized and interpreted with success, emotional speech synthesis has not been as successful. We propose using music as an alternative to speech synthesis. Though it may be possible to use generative music methods, in this first attempt at constructing such a system we use samples of prerecorded music. Such a method allows the system to leverage the many of songs which exist representing a plethora of topics and emotions. Music is an inherently descriptive art form with songs and lyrics capable of conveying complex emotions, topics, and stories. Using audio from the user's musical library as an alternative to speech synthesis allows Shimi to take advantage of music's expressive nature.

The steps for constructing such communicative functionality can be categorized into two parts 1) sentiment and topic detection in speech and 2) tagging portions of songs with particular emotions and topics. In order for Shimi to demonstrate an understanding of human emotion we trained neural networks to classify phrases with particular emotions and topics. Though sentiment and topic labeled corpi exist for NLP, our experience finds that these corpi are not optimal for the interactions we anticipate one to have with an RMC. Therefore, we developed a dataset specifically for our purposes using Twitter similar to the approach used in [8]. Each spoken phrase is then classified by the perceptrons as belonging to one of the six fundamental emotions (happy, sad, surprise, fear, anger, and disgust) and 20 topics ranging from love and relationships to weather, money, and aging.

The second step requires the system to generate an appropriate response. Though automatic mood labeling for entire songs has been performed with marginal success, instantaneous emotion and topic labeling within a song has not been performed and no dataset exists for training. Therefore we hand labeled a library of short audio samples to be used in the system. This allows us to use the quintessential musical representations (based on our own judgement) for each emotion and topic. We are currently working on automating this process, but using hand labeled data enables us to experiment with the HRI component and determine which aspects of the music and lyrics are important for establishing useful communications.

The final interaction consists of the user speaking and Shimi responding by classifying the content and choosing an appropriate sample. For example, a user might say, "I'm so sad my girlfriend broke up with me" and Shimi will detect a sad sentiment and a topic of love. This will result with Shimi playing an audio clip of a sad love song.

4. CONCLUSION

The concept of a robotic musical companion was presented in this work. We provided an overview of several applications for the Shimi robotic platform suggesting that robotics have a place in the many scenarios in which music is experienced. Though several challenges must be addressed and additional research is necessary, novel and entertaining forms of music consumption can arise as a result of creating human-robotic interactions inspired by underlying musical intelligence and ambition. The applications described will continue to be improved upon and further methods for developing an RMC will continue to be explored.

5. REFERENCES

[1] A. Albin, S. Lee, and P. Chordia. Visual anticipation aids in synchronization tasks. *Abstract In Proc. of the 2007 Society for Music Perception and Cognition (SMPC)*, 2011.

[2] C. Breazeal and L. Aryananda. Recognition of affective communicative intent in robot-directed speech. *Autonomous robots*, 12(1):83–104, 2002.

[3] M. Bretan, M. Cicconet, R. Nikolaidis, and G. Weinberg. Developing and composing for a robotic musician. In *Proc. International Computer Music Conference on (ICMC'12)*, Ljubljana, Slovenia, Sept. 2012.

[4] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan. Affect recognition for interactive companions: challenges and design in real world scenarios. *Journal on Multimodal User Interfaces*, 3(1):89–98, 2010.

[5] P. Dahlstedt and P. McBurney. Musical agents: Toward computer-aided music composition using autonomous software agents. *Leonardo*, 39(5):469–470, 2006.

[6] M. E. Davies and M. D. Plumbley. Causal tempo tracking of audio. In *ISMIR*, 2004.

[7] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith. Query by humming: musical information retrieval in an audio database. In *Proceedings of the third ACM international conference on Multimedia*, pages 231–236. ACM, 1995.

[8] G. Hoffman. Dumb robots, smart phones: A case study of music listening companionship. In *RO-MAN, 2012 IEEE*, pages 358–363. IEEE, 2012.

[9] G. Hoffman and G. Weinberg. Shimon: An interactive improvisational robotic marimba player. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 3097–3102, New York, NY, USA, 2010. ACM.

[10] A. Kapur, M. Darling, D. Diakopoulos, J. W. Murphy, J. Hochenbaum, O. Vallis, and C. Bahn. The machine orchestra: An ensemble of human laptop performers and robotic musical instruments. *Computer Music Journal*, 35(4):49–63, 2011.

[11] L. Maes, G.-W. Raes, and T. Rogers. The man and machine robot orchestra at logos. *Computer Music Journal*, 35(4):28–48, 2011.

[12] R. W. Picard. Affective computing, 1995.

[13] J. Solis, K. Taniguchi, T. Ninomiya, T. Yamamoto, and A. Takanishi. Development of waseda flutist robot wf-4riv: Implementation of auditory feedback system. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 3654–3659. IEEE, 2008.