Proceedings of the 17th Linux Audio Conference (LAC-19), CCRMA, Stanford University, USA, March 23–26, 2019

FINDING SHIMI'S VOICE: FOSTERING HUMAN-ROBOT COMMUNICATION WITH MUSIC AND A NVIDIA JETSON TX2

Richard Savery
GTCMT
Georgia Institute of Technology, USA
rsavery3@gatech.edu

Ryan Rose
GTCMT
Georgia Institute of Technology, USA
rrose37@gatech.edu

Gil Weinberg
GTCMT
Georgia Institute of Technology, USA
gilw@gatech.edu

ABSTRACT

We present a novel robotic implementation of an embedded linux system in Shimi, a musical robot companion. We discuss the challenges and benefits of this transition as well as a system and technical overview. We also present a unique approach to robotic gesture generation and a new voice generation system designed for robot audio vocalization of any MIDI file. Our interactive system combines NLP, audio capture and processing, and emotion and contour analysis from human speech input. Shimi ultimately acts as an exploration into how a robot can use music as a driver for human engagement.

1. INTRODUCTION

The field of robotics depends on embedded hardware and software for real-time computational tasks such as kinematics, computer vision, and sensor data processing. For many of these tasks, state-of-the-art performance depends on computationally heavy deep learning techniques. Embedded computing devices have only recently been developed with the GPUs necessary to perform complex deep learning inference in real-time. One such device is the NVIDIA Jetson TX2, an embedded system-on-module that runs Linux on a quad-core ARM processor, and features an 8GB GPU built on NVIDIA's Pascal architecture. This powerful and energy-efficient device greatly expands the capabilities of robots and other embedded applications alike through its ability to run both high CPU and GPU tasks, such as artificial neural networks, deep learning, and signal processing.

This project uses the Jetson TX2 to run a musical robot companion named Shimi (Figure 1). Shimi moves with five degrees of freedom, and can play audio out of two speakers on either side of its head. Additionally, Shimi features a 4-microphone array on its underside. Prior to being run by the Jetson TX2, Shimi was controlled with an Android smartphone and an Arduino Mega.

The purpose of Shimi is to explore novel ways in which humans can communicate with artificial intelligence (AI) agents. Many modern AIs attempt to replicate communicative patterns of humans as closely as possible, using state-of-the-art text-to-speech procedures and complex mechanical operation to try and convince users that they interact with a human-like device, not a computer or a robot. This can quickly lead to the "uncanny valley" psychological phenomenon, where the small differences between an AI and a real human evoke a deeply unsettling feeling. In this project, the authors embrace the non-human robotic identity of Shimi and explore methods of communication using Shimi's limited range of motion and music, in place of verbal language. This is realized through a voice generation system that utilizes deep learning to respond to human speech in an emotionally relevant manner, and a gesture generation system that uses both quantified emotion and Shimi's musical voice to craft robotic body language using Shimi's five degrees of freedom.

Figure 1: The musical robot companion Shimi.

2. RELATED WORK

Prior work on Shimi focused first on utilizing the sensors and computational power of a smartphone to explore the possibilities of personal robotics in a cost-effective way [1]. The research in this study also provided inspiration for life-like gestures, taking cues from animation. Other work on Shimi explored expressing emotion through gesture, informed by observations of human movement and emotion from Darwin [2, 3]. Others have used the Laban Effort System in gesture generation, specifically in low degree of freedom robots such as Shimi [4]. Additionally, speech analysis as input to gesture generation has been used for robot communication in many cases such as *Kismet* [5].

Music as a vector for emotion has been demonstrated in numer-

ous studies, with comprehensive research exploring what emotions can be perceived or induced through music, what musical features encode emotion, and how music expresses or induces emotion [6]. Studies have shown clear correlations between musical features and movement features, suggesting that a single model can be used to express emotion through both music and movement [7]. Additionally, humans demonstrate patterns in movement that is induced from music [8].

3. TECHNICAL DESCRIPTION

3.1. Voice System

3.1.1. Input Analysis

Shimi analyzes incoming audio streams using a combination of natural language processing (NLP) and raw audio analysis. Shimi features a Seeed Studio ReSpeaker Mic Array v2.0 [1], a four-microphone array with on-board processing that combines each microphone stream and denoises the recording, emphasizing voice signals. No additional processing of input signals was added after the ReSpeaker processing, other than down-mixing to a single channel. Using the open-source hotword detection library `Snowboy`[2], Shimi responds to the phrase "Hey Shimi," and begins recording input audio. The Python phrase detection library `speech_recognition`[3] is then used to capture one phrase of raw audio.

Incoming audio is analyzed using the valence arousal model, whereby valence is the measure of the positivity or negativity of an emotion, and arousal is the measure of the energy of an emotion[9]. Raw audio analysis is used to find the arousal level, pitch, intensity and onsets. To do this we utilized `Parselmouth`[4], a Python library built on `Praat`[5]. We created custom metrics to analyze the input level based on analysis of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) data set [10]. RAVDESS includes 7356 audio files by 24 actors, each rated with an emotion independently validated by 10 participants. Our metrics were based on pitch contours and intensity levels found in the recordings. Figure 2 and 3 show analysis of the phrase *the dogs are sitting by the door* from the data set. Our metrics to measure arousal use the variety, level and standard deviation in intensity and the range, contour and standard deviation of pitch.

To measure valence we use the Natural Language Toolkit (`NLTK`) [11], a suite of Python modules for NLP. We calculate valence using a built in naïve bayes classifier trained on the `NLTK` data set of tagged phrases from social media. We also use the `NLTK` library for statement classification.

3.1.2. Shimi's Emotion

Shimi maintains its own emotional state through each communication, tracked through a position in valence and arousal. Valence and arousal are both measured between -1 and 1. The current model gradually shifts the valence level towards that of the user while mirroring the arousal of the user. A negative valence statement from the user will cause Shimi to respond in a sad tone. Following positive statements from the user will gradually move Shimi towards positive responses. When starting Shimi begins with a valence of 0.5, equating to slightly happy.

3.1.3. MIDI Dataset and Phrase Generation

To control Shimi's vocalizations we generate MIDI phrases that then drive the synthesis and audio generation described below and lead the gesture generation. For this purpose we created our own data set of MIDI files tagged with a valence and arousal quadrant. We collected MIDI files from eleven improvisers around the United States. Each was told to record MIDI phrases between 100ms and 6 seconds with each phrase assigned one of the quadrants from the valence/arousal model. They also recorded phrases that they believe represented a question, an answer to a question, a greeting and a farewell. Improvisers were told to record between 50 to 200 samples of each category. To restrict the data each phrase could only contain velocity values at the start of a note and no MIDI data outside pitch, velocity and rhythms were included in training (i.e. no expressive modulations).

As the data set was created by many improvisers we created a second process to confirm the validity of the collected files. This was done through a comparison of the pitch, velocity and contour variation between the new MIDI data set and the RADVESS data set. Figure 3 and Figure 4 present the an example of the variance in the data-set between different emotions (blue is pitch, orange is intensity, placed over a spectrogram). Any MIDI file that varied too far from the features of RADVESS was removed from the data set. Table 1 shows the final amount of files used for Shimi's phrase generation. The RADVESS data set does not include greetings, farewells, questions or answers and due to their limited use in Shimi's interaction we did not post process these phrases.

Table 1: *Shimi Emotional MIDI Data set*

$PhraseType$	MIDI Samples	Post Process
$VA1(Happy)$	895	400
$VA2(Angry)$	1042	621
$VA3(Sad)$	980	567
$VA4(Calm)$	700	385
$Greetings$	655	655
$Farewell$	895	895
$Question$	901	901
$Answer$	778	778

To generate phrases for Shimi vocalizations, we choose to use a data driven generative method. We also considered using the samples recorded by improvisers directly, however we wanted to aggregate the features created by all improvisers and develop a system that allowed limitless variability. Having chosen to use deep learning a relatively simple Long short-term memory, recurrent neural network (LSTM RNN) was implemented in Keras over Tensorflow as has been previously presented [12][13]. This type of neural network is useful for this task as it is sequential and considers parts of its input as it creates output, encouraging the creation of musical phrases. The data set was first transposed into all twelve keys, to avoid a need to identify a key center. Eight different versions of the network were trained, one for each tagged component of the data set. This was done with the goal of a faster run time.

[1] http://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/
[2] https://snowboy.kitt.ai/
[3] https://github.com/Uberi/speech_recognition
[4] https://github.com/YannickJadoul/Parselmouth
[5] http://www.fon.hum.uva.nl/praat/

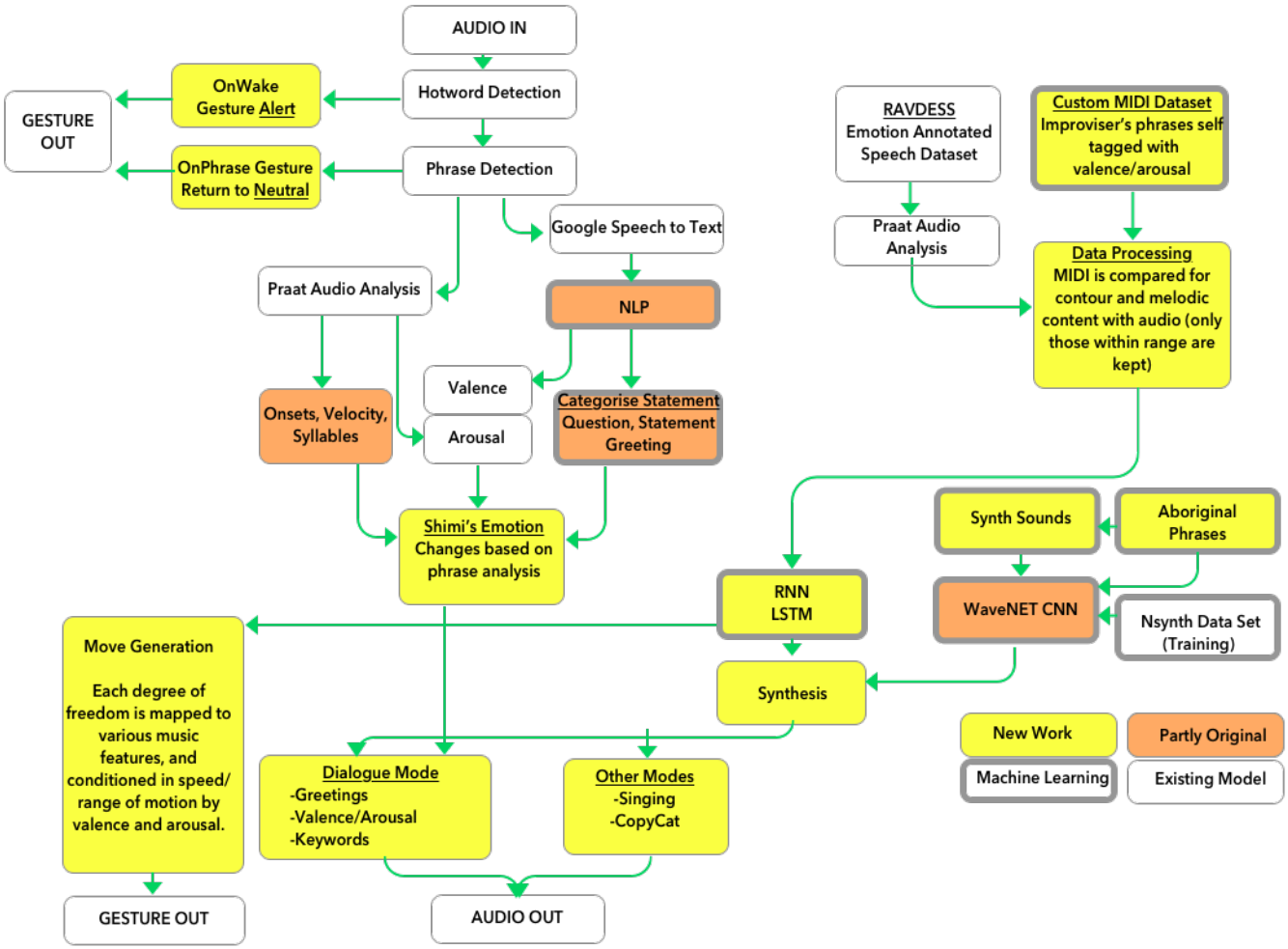

Figure 2: Shimi System Overview.

3.1.4. Audio Creation and Synthesis

MIDI phrases are fed to a new synthesis system created for Shimi. To generate vocalizations that focuses on emotions devoid of all semantic meaning, we chose to construct a new vocabulary. Shimi's vocabulary is built upon phonemes from the Australian Aboriginal language Yuwaalaraay a dialect of the Gamilaraay language. Originally ideas explored real-time implementations of deep learning raw audio synthesis, however it quickly became apparent that this would add unacceptable amount of latency to the system. In our testing even with large compromises in bit rate we were never able to achieve less than a 1 to 5 ratio of processing sound (1 second took 5 seconds to process). Instead of real-time synthesis we compromised by interpolating 28 language samples with four different synthesizer sounds, manually created by the authors. For each sound three different intensity levels were recorded at two different octaves, giving a total of 672 wave samples each 500 ms long. Our final interpolation was done using a modified version of NSynth[14], trained on the NSynth data set. Sounds are played back using a synthesis engine that time stretches and pitch shifts the wave samples to match the incoming MIDI file.

3.2. Gesture System

Much like in human communication, Shimi's gestures are tightly coupled with speech [15]. The voice system produces three outputs: an audio file of Shimi's speech, the MIDI musical representation of the audio, and quantitative measures of Shimi's current emotion. The latter two outputs are the inputs to a rule-based generative gesture system, which controls synchronized playback of gesture with the generated audio.

The first step in gesture generation is musical feature extraction from the MIDI representation of Shimi's speech. Using the Python libraries `pretty_midi`[6] and `music21`[7], musical features such as tempo, range, note contour, key, and rhythmic density are obtained. These features are used to create mappings between Shimi's voice and movement; for instance, pitch contour is used to govern

[6]https://github.com/craffel/pretty-midi
[7]https://github.com/cuthbertLab/music21

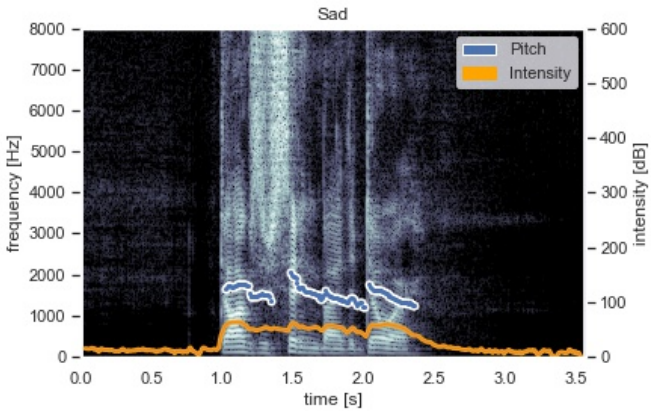

Figure 3: Sad Speech Intensity and Pitch

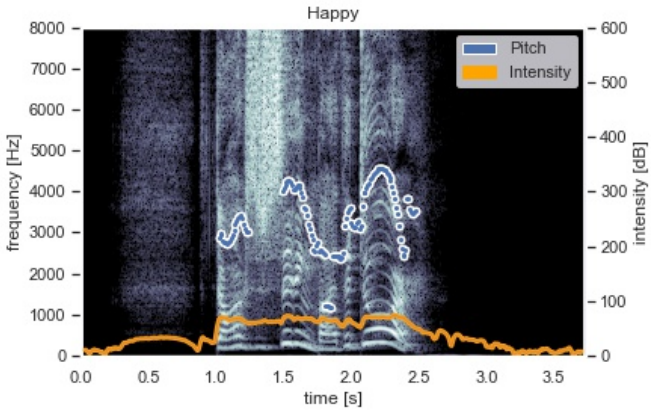

Figure 4: Happy Speech Intensity and Pitch

Shimi's torso forward-and backward movement. Other mappings include beat synchronization across multiple subdivisions of the beat in Shimi's foot, and note onset-based movements in Shimi's up-and-down neck movement. These mappings are based on research investigating correlative features in music and musically-induced movement [8, 7, 16].

The next step uses the emotion state of Shimi to condition Shimi's movement. Emotion is provided to the system in the form of continuous-valued valence and arousal. These values are then used to condition the musical mappings formed previously. In general, arousal is used to restrict or expand range of motion, and valence is used to govern the amount of motion Shimi exhibits, though exact usage varies for each degree of freedom.

In addition to musical and emotional mappings, some degrees of freedom are interdependent. For example, as Shimi's torso moves forward, Shimi's head naturally moves forward and toward the ground. This affects where Shimi is looking, so it is important to consider Shimi's torso position when generating neck up-and-down movement. To accommodate this, the movement paths of Shimi's degrees of freedom are generated sequentially and in full, before being actuated together in synchronization with the audio of Shimi's speech. This is implemented using the built-in `threading` library in Python, with each degree of freedom being associated with one Python thread responsible for sending motor control commands across the duration of the gesture.

The motors used in Shimi are Dynamixel MX-28 actuators produced by Robotis. They feature built-in controllers, allowing for closed-loop control through half-duplex UART serial communication. While the MX-28 motors allow for both reading and writing of position and speed, the half-duplex nature of their communication introduces latency when reading and writing to multiple motors at once, at a resolution high enough for smooth movement. To generate rigorously timed gestures, we do not read Shimi's motors write to them as infrequently as possible. This minimizes any latency inherent in the transmission of data to the motors. For smooth and natural-looking movement, the velocity curve of a gesture is most important. As such, position of Shimi's motors is only ever set when direction of movement changes, and velocity changes are set as frequently as possible without accruing latency. Setting position once and defining the velocity curve allows for control of both when Shimi reaches a certain position, and how Shimi gets there.

Gestures, then, are defined as sequences of movements to a position over a specified time. To facilitate programmatic gesture generation, a collection of velocity curves have been implemented to provide styles of movement. The simplest is a constant velocity, where velocity is the distance of the movement over its duration (Figure 5). This style looks the most stereotypically "robotic", as the motors can accelerate from rest to max velocity much faster than a human can.

Previous work on Shimi introduced a velocity curve that features a constant acceleration until the midpoint of the gesture, then a constant deceleration [1]. This works particularly well for single movement or broad gestures, and looks the most realistic when compared with human motion (Figure 5).

In the context of a multi-move gesture, however, accelerating and decelerating every movement becomes unnatural, as multi-movement human gestures do not come to rest bewteen each move. Thus, a constant acceleration (or deceleration) and constant velocity curve can cap both ends of a gesture. An example of the acceleration variety is shown in Figure 5.

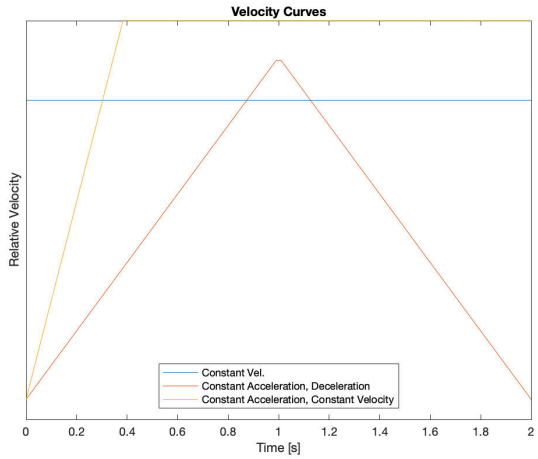

Figure 5: Graphs of the velocity curves used for Shimi movements.

In addition to the movement sequencing method of gesture generation, a different method of recording and playing back gestures is being explored. This method requires physically moving Shimi's

limbs in a desired gesture while the motors continuously record position and speed as fast as possible. After recording, the captured positions and speeds can be used to actuate the gesture on Shimi on demand, resulting in a highly detailed and smooth gesture. While this method results in the most nuanced and expressive gestures, there are difficulties in playing back recorded gestures accurately in time with the way they were recorded. The time taken to read a motor's position and speed varies, resulting in playback that is not aligned with the recording. This timing behavior makes synchronization with speech, which is a necessity for Shimi, very difficult. More research on ways to align these types of gestures with audio is being explored.

4. APPLICATIONS AND FUTURE WORK

This work has described Shimi's ability to generate musical and gestural responses to human speech input that attempts to replicate the emotion conveyed in a spoken phrase. These short form interactions provide insight into how robots can express emotion and communicate with music. A next step in communication will be seeing how accurately Shimi can imitate a phrase, both vocally and, more importantly, emotionally. We are also interested in expanding Shimi's musical phrases to include more languages and improvisers of different origins.

Shimi originated as a musically-intelligent speaker dock, and the work presented here can extend to more musical applications as well. One possibility is as a nuanced music recommendation system. In this system a human would ask Shimi if they would like a song, and Shimi would reply with a vocalization and gesture demonstrating an opinion of that song. This way of expressing opinion can be much more detailed than the thumbs up/thumbs down of many music service providers today. Another engaging musical experience furthers a previous goal of the Shimi project: to enjoy one's music alongside a human listener. Now that Shimi has a voice, the ability to dance along with one's music can incorporate singing along as well. This could also lead to Shimi as a robotic performer, listening to human performers and improvising alongside as a vocalist.

5. ACKNOWLEDGMENTS

Thanks to Matthew Kaufer and Yashveer Singh.

6. REFERENCES

[1] Guy Hoffman, "Dumb robots, smart phones: A case study of music listening companionship," in2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, Paris, sep 2012, pp. 358–363, IEEE

[2] Charles Darwin and Phillip Prodger, The expression of the emotions in man and animals, Oxford University Press, USA, 1998.

[3] Mason Bretan, Guy Hoffman, and Gil Weinberg, "Emotionally expressive dynamic physical behaviors in robots,"International Journal of Human-Computer Studies, vol. 78, pp. 1–16,jun 2015

[4] Heather Knight, "Expressive motion for low degree-of-freedom robots," Jul 2018. https://pdfs.semanticscholar.org/69e2/dc8eab578131a536396a812a2306b6796477.pdf

[5] Cynthia Breazeal and Lijin Aryananda, "Recognition of Affective Communicative Intent in Robot-Directed Speech," p. 20, 2002. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.7147rep=rep1type=pdf

[6] Patrik N. Juslin and John A. Sloboda, "15 - music and emotion," in The Psychology of Music (Third Edition), Diana Deutsch, Ed., pp. 583 – 645. Academic Press, third edition, 2013

[7] Beau Sievers, Larry Polansky, Michael Casey, and Thalia Wheatley, "Music and movement share a dynamic structure that supports universal expressions of emotion,"Proceedings of the National Academy of Sciences, vol. 110, no. 1, pp. 70–75, jan 2013

[8] Petri Toiviainen, Geoff Luck, and Marc R Thompson, "Embodied Meter: Hierarchical Eigenmodes in Music-Induced Movement,"Music Perception: An Interdisciplinary Journal, vol.28, no. 1, pp. 59–70, sep 2010.

[9] James A. Russell, "A circumplex model of affect, "Journal of Personality and Social Psychology, 1980. https://www.researchgate.net/publication/235361517_A_Circumplex_Model_of_Affect

[10] Steven R Livingstone and Frank A Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, "PLOS ONE, vol. 13,no. 5, pp. 1–35, 2018

[11] Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing with Python, O'Reilly Media, Inc., 1st edition, 2009.

[12] Andrej Karpathy, "The Unreasonable Effectiveness of Recurrent Neural Networks,"Web Page, 2015

[13] Richard Savery and Gil Weinberg, "Shimon the Robot Film Composer and DeepScore", "Proceedings of Computer Simulation of Musical Creativity, August 2018, Dublin, Ireland. http://galapagos.ucd.ie/wiki/pub/OpenAccess/CSMC/Savery.pdf

[14] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi, "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders, "CoRR, vol. abs/1704.0, 2017.

[15] David McNeill, How language began: Gesture and speech in human evolution, Cambridge University Press, 2012.

[16] Birgitta Burger, Suvi Saarikallio, Geoff Luck, Marc R. Thompson, and Petri Toiviainen, "Relationships Between Perceived Emotions in Music and Music-induced Movement,"Music Perception: An Interdisciplinary Journal, vol. 30, no. 5, pp.517–533, June 2013.