Automatic Accompaniment of Poem Recitation

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Avneesh Sarwate Georgia Institute of Technology Center For Music Technology avneeshsarwate@gmail.com |  |
|  |

ABSTRACT

This project is an installation piece that attempts to wed together poetry and music for musical novices. It is an automatic accompaniment system that will generate music as a user reads a predetermined poem into a microphone, both harmonizing their voice and generating a bassline. The system captures the arm gestures of the user to parameterize audio effects on the harmonized voice and bassline. The melodic content is algorithmically generated via a pre-determined Markov model.

KEYWORDS

ACM proceedings, text tagging

1 CONCEPT

My project is an automatic accompaniment system that will generate music as a user reads lines of a pre-selected poem. The lines of the poem will be displayed on a screen, and the user is encouraged to dramatically orate the text. The system will automatically harmonize the voice of the user to a generated chord progression, and will also generate a bassline based on those chords and the user’s voice input. The system will track the arm movements of the user to control different effects on the voice.

My system is intended to be a performative installation piece. Because it utilizes a webcam as input it would be installed in a fixed location with a solid colored background and consistent lighting. The user would face a screen that displays the text of the poem to be read, one line at a time. A microphone would be placed in front of the screen for the user to speak into. The screen would ideally be at chest level and angled upwards, like a podium. This would allow the user to read the words and still look up and make eye contact to an audience in front of them.

2 CONTEXT

The core motivation for this project was to create an accessible piece of interactive music around the poem Paradise Lost. I am involved in the production of a musical theater rendition of the poem, and wanted to create a standalone work that could be used as a teaser for the production.

The core conceptual motivation for this installation came from the performance of poetic epics. I wanted to capture the experience of a powerful orator performing enthusiastically such that, in the eyes of the audience, the performance becomes almost supernaturally powerful. I wanted to give this same type of “superhuman” oration power to a novice musician. It was this line of thinking that led to decision of using the processed human voice, rather than a plethora of generated synthetic instruments, as the focal point of the sonic output.

The system also draws its inspiration from the work of Jacob Collier. Collier is an American Jazz musician who is known for playing a vocal-harmonizing keyboard [3]. His performances involve him singing melodies, while his keyboard pitches his voice to the notes selected on the keyboard, while preserving the vowels and consonants so as to still leave the words intelligible. This type of harmonization fits well into the metaphor of “superhuman oration” as described above.

Again, in line with the oration theme, I did not want to incorporate any external controllers, and decided that the only inputs to the system would be the voice and the movements of the user as they are performing. In particular, I decided to focus on using the arm movements of the user as the primary control input, as these were relatively natural gestures to make while reading a written text out loud.

3 IMPLEMENTATION

3.1 Harmonizer

The main component of the system is the vocal harmonizer. This component takes in audio input from a microphone, and MIDI notes from the chord generator (described later), and re-pitches the voice of the user to the input pitches specified by the MIDI notes. The harmonizer is built using the poly~ object, allowing multiple pitches to be played at once. The subpatches of the poly~ object use the retune~ object to implement pitch shifting.

One of the vocal effects that can be controlled by the arms is sub-octave pitch shifting, and due to latency considerations, this is implemented within the poly~ subpatch as well. The naive implementation would have been to run the output of the poly~ into another retune~ object and the mix it back with the original signal (as shown in Figure 1). 

**Figure 1:** Naïve sub-octave implementation

However, this approach leads to the sub-octave signal being 1 signal block behind the standard pitch signal, and creates unwanted modulation. Instead, the mixing of standard pitch and sub-octave is done within the poly~ sub patch itself (as shown in Figure 2).



**Figure 2:** Parallelized sub-octave implementaiton

3.2 Chord Selection

The chords that determined what notes were played are determined by a simple first order Markov model with 5 states, with each state corresponding to a pair of chord. A user is given 3 seconds to read a line, with a progress bar showing how much time was left (figure 3). One step of the Markov model is executed per line, with only one chord being played per line.

The decision of what chord to play out of the pair of chords in a state is made using information extracted from the line of the poem being read. Each pair of chords consists of one diatonic 7th chord, and one diminished chord, rooted at the same note. The choice of whether to play the diatonic or diminished 7th is based on the valence of the sentence, valence being the overall positivity or negativity of a piece of text [4]. If the valence is greater than zero, the dominant seventh is selected, and if it is less than zero, the diminished chord is selected. If it is zero, the valence of the most recent non-zero-valence line is used to determine the chord type. The sentiment analysis is done using the VADER model, which is distributed in the Natural Language Toolkit package for Python [1,2].

3.3 Arm Position Estimation

The arm movements of the user control 4 different parameters, the specifics of which will be explained in the next section. The 4 input dimensions are the height and distance-from-body of the left and right arms, respectively. These positions are estimated using only a webcam.

The estimation is done as follows. First, the user is assumed to be standing in the center of the frame. A snapshot of the “background” including the user is taken. Then, the background is compared to the live-webcam frames. To allow for independent feature extraction of left and right arms, the background frame and camera frame are split vertically down the center. For each arm, the “difference” frame is calculated - this is a 2D matrix where diffMatrix[i,j] corresponds to the difference between the colors of the pixels at backgroubd[i,j] and camera[i,j]. The color distance function used was developed by CompuPhase [5]. Next, I calculated the center of mass of the difference frame. X and Y positions of the center of mass, normalized to 1 by the dimensions of the difference matrix, are used as the output values for arm position.

There two more steps of noise-reduction and adaptive scaling implemented on top of this to generate the final output values, but their description is left to comments within the source code for better context.

3.4 Bassline Generation

The system automatically generates a bassline that plays underneath the user’s harmonized voice. The notes of the bassline are chosen at random from the pitches of the currently playing chord. The rhythm of the bassline is generated from the user’s speech pattern. The system performs onset detection on the user’s voice, and records the timestamps of these onsets. The onset pattern generated while the user speaks a line of the poem is quantized to 16th notes (with the duration of the line assumed to be 1 measure in 4/4 time), and that pattern and is used as the rhythm of the bassline in the next line.

3.5 Harmonizer

The 4 dimensions of arm movement control 4 different effects. The horizontal movement of the left arm controls the mix of the sub-octaved vocal harmony. Having the arm closer to the body produces less sub-octaving, and having it further increases it. The vertical position of the left arm controls the amount of echo. A higher arm corresponds to more echo and a lower arm to less. The horizontal position of the right arm controls the amount of additional hits are randomly added to the rhythm of the bassline. Moving the arm closer to the body produces fewer additional hits and moving it further produces more. The vertical position of the right arm determines the mix of a grain delay on the bass track. A higher arm corresponds to a higher mix, and vice versa.

4 STRENGTHS AND WEAKNESSES

I believe that the aesthetic choices of the system, particularly the choice of input gestures and their corresponding effects, are quite solid. The types of effects chosen fit well within the mood of paradise lost, with sub-octave, granulated textures fitting well with the theme of Hell and with smooth, spacey major chords corresponding well to heavenly ethereality. The gestures (if implemented well), seem to be a natural way to control the real time, continuous input effects - the values can be both slowly changed and suddenly shifted and the perception of the change is immediate.

However, the variability of the feature extraction, as well as the relatively low audio quality of the pitch shifting, do not make for the most pleasant experience. The variability of the output based on the lighting conditions and the distance of the user to the camera makes it difficult to determine a proper value with which to scale the arm dimensions, leading to a substandard control. A proper installation would require relatively careful calibration of these parameters, and perhaps even a more sophisticated strategy that adaptively rescales the output based on the user’s movements. Also, the quality of the pitch shifting is lacking, particularly for the sub-octave effect, where it produces atonal chords and sometimes clicking artifacts.

The system is simple, but as an installation, it is unclear whether it will be transparent if no instructions are provided. Because a “snapshot” of the user and background must be taken to calibrate the arm detection, the system could be installed as a sort of “booth” with an operator that could provide instruction. However, this necessity for calibration greatly detracts from the smoothness of the experience.

Another question is how users will interact with such an installation in a public setting. Given that public speaking is a notoriously common fear, it is unclear how many people would be willing to step up and give a dramatic reading to a room full of people.

5 FUTURE WORK

Future work for this system would include developing more robust methods of hand tracking. There are still a handful of adaptive thresholding tricks that I did not have time to implement. The initial hope had been to find a way to implement predictable arm tracking without the need for special hardware, but a potential avenue for improvement is to re-implement this with an XBox Kinect.

Another line of future work could be to modify this project to be used as part of a stage set up during a theater piece, allowing an actor to deliver a monologue and generate his or her own accompaniment. This usage would allow for more time for detailed calibration, and would avoid the problem of users being hesitant to interact with the system.

REFERENCES

|  |  |
| --- | --- |
| [1] | Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Tool Kit. In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia, July 2002. |
| [2] | Hutto, C.J. & Gilbert, E.E. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014. |
| [3] | "Jacob Collier." Arts at MIT. Accessed December 14, 2017. https://arts.mit.edu/artists/jacob-collier/#about-the-residency. |
| [4] | Mohammad, Saif M. 2015. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Emotion measurement: 201-238. |
| [5] | Riemersma, Thiadmer. 2012. "Colour metric." Accessed December 14, 2017. https://www.compuphase.com/cmetric.htm. |