FutureGrab: A wearable synthesizer using vowel formants

Yoonchang Han
Music and Audio Research Group
Seoul National University
Seoul, Republic of Korea
yoonchanghan@snu.ac.kr

Jinsoo Na
Music and Audio Research Group
Seoul National University
Seoul, Republic of Korea
evencra7@snu.ac.kr

Kyogu Lee
Music and Audio Research Group
Seoul National University
Seoul, Republic of Korea
kglee@snu.ac.kr

ABSTRACT

FutureGrab is a new wearable musical instrument for live performance that is highly intuitive while still generating an interesting sound by subtractive synthesis. Its sound effects resemble the human vowel pronunciation, which were mapped to hand gestures that are similar to the mouth shape of human to pronounce corresponding vowel. FutureGrab also provides all necessary features for a lead musical instrument such as pitch control, trigger, glissando and key adjustment. In addition, pitch indicator was added to give visual feedback to the performer, which can reduce the mistakes during live performances. This paper describes the motivation, system design, mapping strategy and implementation of FutureGrab, and evaluates the overall experience.

Keywords

Wearable musical instrument, Pure Data, gestural synthesis, formant synthesis, data-glove, visual feedback, subtractive synthesis

1. INTRODUCTION

The rapid development of digital technology within the past few decades has extended a range of sound synthesis possibility. A lot of new kinds of synthesizers appear on the market everyday, and the ever decreasing cost of various electronic parts and simple yet powerful tools such as Pure Data[1] and Arduino[2] have enabled even individual users to design and make musical devices. One of the most widely used electronic musical devices nowadays is synthesizer. It generates sound by using various combinations of different types oscillators, mixer, filters and envelopes. However, it is often too hard for beginners to use, because altering lots of synthesis parameters to design intended sound is difficult without long-term practice. This has therefore provided the motivation for making an improved control interface of a musical instrument for live performances that is highly intuitive while still generating an interesting sound with a wide range of timbre. In order to make an intuitive interface, first we needed to find a link between sound synthesis and usability. High usability usually comes from familiarity with a tool, and one of the most familiar sound syntheses we found in our daily lives was our voice. The process of human voice generation resembles a subtractive synthesizer, because the way vocal tract shapes the sound generated from the vibration of the vocal fold is basically identical to the process of filtering the source in subtractive synthesis.

The main idea came from the fact that people are well aware of what mouth shape is needed to pronounce certain vowel. This has led us to map filtering variables for vowels to hand gestures that resemble mouth shape of human. In addition, a pitch indicator was added to give visual feedback on the current pitch to minimize the mistakes during live performances, because a minor problem about pitch accuracy arose during the evaluation process of the initial version of FutureGrab. It is known that real-time visual feedback on the pitch is a useful tool for increasing intonation accuracy of the musical instrument [1].

2. BACKGROUND

2.1 Related Work

The first of its kind, and probably the most well known electronic musical instrument using hand gesture is the Theremin. It consists of two antennas to detect position of performers hand to control oscillator frequency with one hand and amplitude with the other [2]. One of the first works that maps formant to data-glove gesture is Glove-talk [3]. Glove-talk is a speech synthesizer, and it enables users to pronounce words using 66 root words. Glove-talk allows a wide range of freedom in terms of pronunciation, but each hand gesture needs to be remembered to synthesize words. The primary aim of FutureGrab is not speech synthesis, but a musically interesting sound. Thus it is fair to say that the only similarity among Glove-talk and other speech synthesis data-glove such as ForTouch [4], GRASSP [5] and Future Grab is that formants were mapped to the data-glove.

There are a number of previous works using hand gestures for musical purpose, such as Cyber Composer [6] and SoundGrasp [7]. Cyber Composer allows users to control the melody flow generation, pitch and volume, and SoundGrasp does sampling, looping and adding sound effect on it. The required gestures of these data-gloves are simpler than Glove-talk. However, the link between the mapped gestures and the actual synthesized or manipulated sound is weak. Hence, users still need a fair amount of time to practice prior to actual live performance. One of the most well known vowel-like sound effects is a Wah-wah

[1] http://puredata.info/
[2] http://arduino.cc/

NIME'12, May 21-23, 2012, University of Michigan, Ann Arbor.

pedal. The sound is produced mostly by a foot-controlled signal processor containing a single band-pass filter with a variable center frequency [8]. Although using a single filter is not sufficient to generate accurate vowel, it is still popular as it can be heard as somewhat similar to a 'wah' sound, which is musically interesting. The main advantage of wearable musical instruments is natural musical expression. Detailed strategies for mapping gesture variables naturally to sound synthesis is explained in [9], and factors of design and user experience evaluation of wearable expression were explained in [10].

2.2 Vowel and Formant

Formants are spectral peaks of the sound spectrum envelope of the voice [11]. The human voice is generated by the vibration of the vocal fold, which is a spectrally rich acoustic excitation. The generated sound source is then shaped by the vocal tract, an adjustable acoustic filter in our body to pronounce certain word [12]. Each vowel almost always has several certain formants, although there might be a slight difference between people. Discrimination of human vowels chiefly relies on the frequency relationship of the first two peaks of the vowel's spectral envelope [13][14][15]. Therefore, FutureGrab makes use of the first two formants only for filtering, which is sufficient to create vowel-like effects. Figure 1 shows the first two formant frequencies of the vowels /i/ and /u/, respectively.

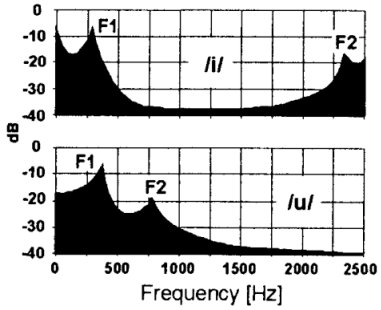

Figure 1. First two formant of /i/ and /u/ in spectrogram envelope (adapted from Frank and Henning [14])

3. SYSTEM DESIGN

3.1 System Architecture

The system overview of FutureGrab is shown below in Figure 2.

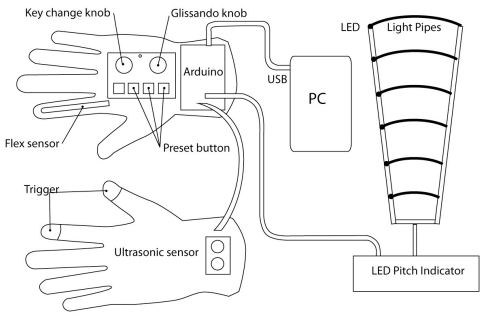

Figure 2. The system overview of FutureGrab

The primary interface consists of two gloves manipulated by the performer. The gloves are connected to each other and the right-hand glove, which contains an Arduino chip, was connected to a computer that runs Pure Data. Arduino is a small microcontroller that can handle input from a variety of sensors as well as controlling the output of actuators. All input sensors, buttons, and output LEDs, were connected to the Arduino. Pure Data is a real-time visual programming language for interactive computer music, and is used for sound synthesis, filtering and adding various effects on a signal. Communication between Arduino and Puredata was done by a custom version of Firmata[3], which was modified to handle digital ultrasonic sensors.

3.2 Mapping Strategy

Filtering frequencies to make vowel-like formants were mapped to how much the hand was clenched, which was measured by a flex sensor on the right index finger. Vowels that requires a nearly closed mouth, such as /u/ and /e/, are mapped to a clenched hand, and vowels that require a fully opened mouth shape, such as /ɑ/ and /e/, were mapped to am opened hand. The user can select which vowel to use by using preset buttons. Figure 2 is a comparison between human mouth shape to pronounce vowel /ɑ/ and /u/, and required hand gesture.

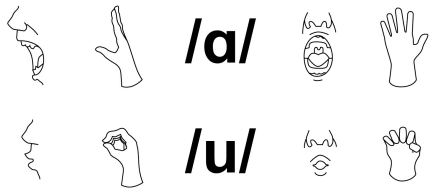

Figure 3. Mouth shapes were mapped to hand gestures and synthesized sound

Formants from a clenched hand to an opened hand change gradually, resulting in vowel sounds changing naturally from one to another. For instance, if the preset is /u/ to /ɑ/ and the hand is only half-clenched, the resulting sound will be close to /o/ because the actual formants of '/u/' is actually somewhere between /u/ and /ɑ/. This smooth change between two vowels was mapped to clenching the right hand. The pitch of the source signal was mapped to the distance between the left palm and the ground, which is measured by an ultrasonic wave sensor. This sensor measure the distance from the sensor to anything that blocks the ultrasonic wave, thus can be actually used against any flat surface such as a desk, wall, or even the performers chest. The target object can be chosen flexibly depending on the environment or preference of the performer. The trigger function was mapped to a pinching gesture, in which the user puts his left index finger on his left thumb. Figure 4 shows FutureGrab gestures for pitch control, trigger and sliding filtering frequency.

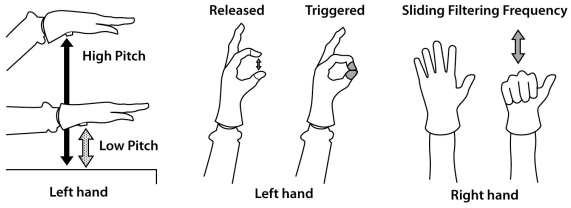

Figure 4. Mapping between hand gestures and functions of FutureGrab

FutureGrab is a monophonic synthesizer, thus it is suitable to use as a lead instrument to play melody parts of the music. In order to match the accompaniment in different keys, a key change function was added and controlled by the knob at the back of the right glove. Also, the C minor 'blues scale'[4] was set

[3] A generic protocol for communicating with microcontrollers from software on a host computer. `http://firmata.org/`

[4] Minor pentatonic scale plus the ♯4th or ♭5th degree.

as the default, because it is the most widely used scale in modern popular music.

Although FutureGrab is simpler than typical synthesizers, it still provides most of the main functionalities of existing subtractive synthesizers. For instance, the ultrasonic distance measure with the trigger substitutes a keyboard; clenching and opening hand gestures can be thought as adjusting a filter envelope; and the source signal from Pure Data corresponds to an oscillator. In addition, a glissando[5] function was added to make a gradual pitch change, which is common in monophonic synthesizers, which was controlled by the knob attached next to the key change knob.

3.3 Formant Synthesis

FutureGrab mimics the process of human voice generation, which includes a source generation followed by filtering. The source signal was generated by Pure Data and the resulting sound was then shaped by two band-pass filters with variable center frequencies (f_{c1}, f_{c2}) to create both first (F1) and second formant (F2). Unlike the wah-wah filter, the popular vowel-like sound effect, which uses a single band-pass or low-pass filter, using two filters gives the freedom to imitate any vowel by combining two formants. Figure 5 shows how vowel formant-like spectrum was created using two band-pass filters.

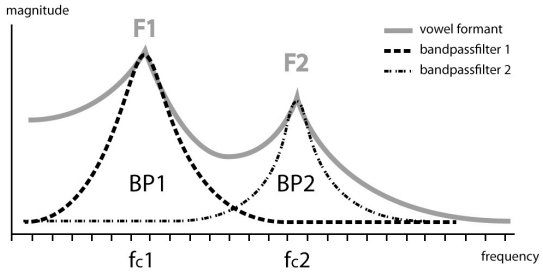

Figure 5. Using two band-pass filters to shape source signal into human formant like spectral shape.

The required band-pass center frequencies for the first two formants of each vowel are shown below.

Table 1. First (F_1) and second (F_2) formant of vowel used in FutureGrab. (Data extracted from Peterson and Barney [15])

Vowel (IPA[6])	F_1	F_2
/u/	300	870
/ɑ/	730	1090
/i/	270	2290
/e/	660	1720

The pitch was adjusted by changing the fundamental frequency of the source signals and designed to cover up to three octaves. When the key-change knob is adjusted, the fundamental frequency moves to the pitch of the requested key.

4. IMPLEMENTATION

The implementation process of FutureGrab is mostly about signal processing in Pure Data and how electronic parts were connected, which are platform-specific technical details and not appropriate to this paper. Thus, we aim to explain the general overview of the implementation process in this section, which can be freely applied to any system regardless of the platform.

4.1 Software

Three sawtooth waves were used to create a harmonically rich sound source. In addition, the gain for each oscillator was set slightly different similar to existing subtractive synthesizers and a little bit of noise was added in order to maximize the richness of the sound. The measured distance from the ultrasonic sensor was quantized to make a discrete pitch scale. Also, the amount of glissando was adjusted by the speed of change in the fundamental frequency on the note change. When the speed of change between the current note and the next note was very fast, it sounds like two discrete notes, and when the change is slow it makes a more continuous sound. This degree of continuousness was programmed to affect the changes in the LED lights of the pitch indicator as well, so that the performers and audience can see the effect of the glissando visually. The center frequency of the band-pass filters were controlled by a flex sensor and designed to slide the formant between two different vowels. Q factors[7] of the band-pass filters were empirically chosen as 30 as we considered that this value produced the best resulting sound, but it can be freely adjusted depending on the preference of the performer.

4.2 Hardware

An Arduino was used for obtaining sensor values and controlling actuators as mentioned in the system architecture. Every electronic part were soldered on a circuit board and inserted between the inner and outer skin of the glove to make it hidden. The distance between the left palm and the ground for pitch control was measured ultrasonic sensor, and the trigger was made by covering the thumb and index finger with a copper wire such that the circuit is coupled when the user does a pinching gesture. Since the Arduino used in this project did not have sufficient digital ports, we designed an analog circuit to handle the preset buttons. This was done by measuring the voltage difference caused by parallel resistors. The flex sensor was directly connected to the analog port, as it returns analog values. Knobs for key adjustment and glissando effects were implemented using a potentiometer. The pitch indicator was made of copper pipes with vertically attached high intensive blue LEDs. Horizontal lights were created using LEDs and optical pipes. In addition, a copper pipe was used as a ground. Figure 6 below is a picture of FutureGrab and pitch indicator.

Figure 6. Picture of FutureGrab and the pitch indicator

5. EVALUATION

To explore the usability and intuitiveness of FutureGrab as a musical instrument, evaluation was done through a questionnaire. This section includes the results of testing and explanation about what was improved after the evaluation. The participants were composed of 12 graduate students from Seoul National University who were not familiar with FutureGrab.

[5] Musical term that means gliding from one pitch to another.

[6] IPA (International Phonetic Alphabet) is standardized representation of the sounds of spoken language.

[7] Quality factor is a bandwidth relative to its center frequency.

All participants were asked to try every feature of FutureGrab without any instructions. The first question of the questionnaire was, "How relative did you find the link between the synthesized sound and the gestures?" which was answered using a rating system between 1 and 5, where 1 indicates that they could not find any relationship between them. The second question, a multiple-choice question, was "What was the hardest feature to use?"

As a result, the users found the link between gestures and synthesized sound easily. The average rating was 4.25, and there were no ratings below 3. Although the first question showed that FutureGrab is highly intuitive, problems with pitch accuracy was pointed out. 7 out of 12 participants chose pitch control as the hardest feature to use in the second question. When the evaluation was carried out, the pitch was shown as Solfège[8] with a key status on the computer screen. This was sufficient to play the correct pitch, but the problem was that it was confusing when changing the pitch quickly. Hence, after the evaluation, we decided to make a large size pitch indicator that displays the pitch using vertical positions of the light.

6. DISCUSSION

There was a chance to use FutureGrab in live performance, and we could check the reaction of the audience and the potential of FutureGrab as a popular musical instrument. As a result, the FutureGrab live performance was fairly successful. Especially, the fact that everyone can see what the musician is actually doing to generate sounds, unlike other instruments that are barely visible to the audience, caught the interest of the audience. Unexpectedly, one of the things that impressed the audience the most was the LED pitch indicator. It was mainly designed for pitch accuracy of the performance, but its fascinating high intensive blue LED that follows the melody of the music drew loud cheers from the audience.

7. CONCLUSIONS AND FUTURE WORK

We have described a mapping strategy and the development process of FutureGrab, which is a novel wearable instrument particularly designed for live performances. Its sound effect was inspired by the vowel pronunciation of humans, and was possible to achieve high intuitiveness by mapping familiar human mouth shapes to hand gestures and synthesized sound at the same time. FutureGrab has a relatively simple structure, but still provides most of the main functionalities of the existing subtractive synthesizer. FutureGrab was originally designed from the perspective of the performer mainly for the musical performance. However, several strong advantages in terms of visual presentation have emerged during the development process. We found that showing hand gestures and the LED pitch indicator to the audience greatly improves the visual presentation in actual live performances.

In the future, we plan to make the next version of FutureGrab portable, independent, and durable as possible. The current version of FutureGrab runs Pure Data on the computer for signal processing, which means that it is necessary to bring extra equipment such as a laptop, adaptor, and lots of wires. We are planning to put a small-sized computer inside of the pitch indicator so that FutureGrab can be a standalone musical instrument, which will greatly increase the portability. Also, using wireless protocols such as Bluetooth or ZigBee for data communication between gloves would allow us to remove the limitation of movement of the performer as well as preventing potential problems that might be caused by the wire connections. Ultimately, we wish to make a final version of FutureGrab as a totally independent synthesizer with an embedded DSP chip and microcontroller so that it works without Pure Data and OS for the highest possible stability and cost-effective mass production.

[8] Solfège is a relative note name such as do, re and mi.

8. ACKNOWLEDGEMENTS

We would like to thank the members of the Music and Audio Research Group who helped further develop the ideas, and the participants who volunteered to take part in the user survey. Our special thanks go to Yongtae Hwang who shared his expertise in hardware design with us.

9. REFERENCES

[1] D. A. Smith and J. D. Lehman. The Effectiveness of Real-time Visual Feedback to Improve Seventh and Eighth Grade Saxophone and Trombone Students' Intonation Accuracy, Purdue University, 2006.

[2] A. Glinsky, *Theremin: Ether Music and Espionage*. Urbana, Illinois: University of Illinois Press, 2000, 24-25.

[3] S. Fels and G. E. Hinton. Glove Talk: A Neural Network Interface Between a Data-Glove and a Speech Synthesizer. *IEEE Trans. on Neural Network. 4*, 1 (Jan. 1993), 2-8.

[4] S. Fels, R. Pitchard and A.Lenters. For Touch: A Wearable Digital Ventriloquized Actor. In *Procedings of the International Conference on New Interfaces for Musical Expression (NIME),* Pittsburgh, PA, USA, 2009, 274-275.

[5] S. Fels, R. Pitchard and A.Lenters. GRASSP: Gesturally-Realized Audio, Speech and Song Performance. In *Procedings of the International Conference on New Interfaces for Musical Expression (NIME06),* Paris, France, 2006, 272-276.

[6] H. Ip, K. Law and B. Kwong. Cyber Composer: Hand Gesture-Driven Intelligent Music Composition and Generation. In *Proceedings of the 11th International Multimedia Modelling Conference (MMM*), 2005.

[7] T. Mitchell and I. Heap. SoundGrasp: A Gestural Interface for the Performance of Live Music. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME).* Oslo, Norway, 2011, 465-468.

[8] U. Zölzer and X. Amatriain. *DAFX: Digital Audio Effects,* John Wiley and Sons, NJ, USA, 2006.

[9] M. Wanderley and P. Depalle. Gestural Control of Sound Synthesis. *Proc. of IEEE. 92,* 4 (Nov. 2004), 632-644.

[10] J. Nugroho and K. Beilharz. Understanding and Eveluating User Centred Design in Wearable Expressions. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME).* Sydney, Austrailia, 2010, 327-330.

[11] G. Fant. Acoustic Theory of Speech Production. Mouton & Co, The Hague, Netherlands, 1960.

[12] H. Pulakka. Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography. Helsinki University of Technology, Helsinki, Finland

[13] P. Delattre, A. Liberman, F. Cooper and L. Gerstman. An Experimental Study of the Determinants of Vowel color; Observation on One- and Two-Formant Vowels Synthesized From Spectrographic Patterns. *Word, 8,* 1952, 195-210

[14] W. O. Frank and S. Henning. Orderly Cortical Representation of Vowels Based on Formant Interaction, In *Proceedings of the National Academy of Science of the USA (PNAS), 94,* (Aug. 1997), 9440-9444

[15] G. E. Peterson and H. L. Barney. Control Method Used in a Study of the Vowels. In *Journal of the Acoustical Society of America (JASA), 24, 2,* (Mar. 1952), 175-184