Music in the Air with Leap Motion Controller

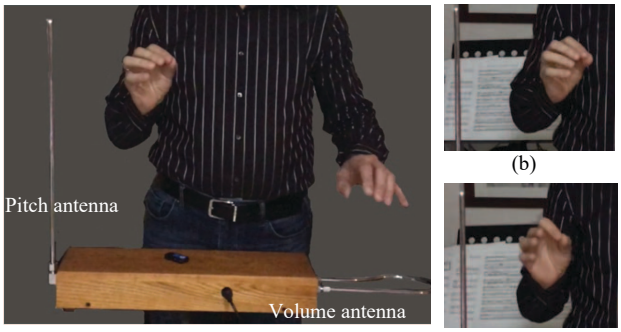
Alexei Sourin
School of Computer Science and Engineering
Nanyang Technological University
Singapore
e-mail: assourin@ntu.edu.sg

***Abstract*—Not many people know about the first electronic musical instrument—the theremin—and can play it. The idea of this instrument is very groundbreaking: it is played without physical contact with it and in the same way as we sing but by using hands in place of our vocal cords. In this paper we consider how to implement the theremin with a computer using very different physical principles of optical hand tracking and by adding advantages of visual interfaces. The goal of this research is to eventually fulfill the dream of the inventor to make the theremin a musical instrument for everyone and to prove that everyone can play music.**

Keywords-multimodal interaction and rendering, art and heritage in cyberspace, theremin

I. MUSIC FROM THE AIR

One hundred years ago, the first electronic musical instrument—*theremin*—was invented by Russian scientist Leon Theremin (Lev Sergeyevich Termen). It is played by the hands of the performer without physical contact with the instrument so that the position of one hand in the air controls the pitch of the sound while the position of another hand – its volume. Moog Etherwave theremin displayed in the Fig. 1 has a reliable pitch range from 1.5 octaves below middle C to 2.5 octaves above it.



(a) (b) (c)

Figure 1. (a) Playing the theremin, (b) closed fingers position, (c) extended fingers position.

The instrument design is based on using two metal antennas to sense how far the hands are located from them. In simple terms, the hands act as grounded plates of variable frequency capacitors in two inductance-capacitance circuits. Motion of the hands dynamically changes frequencies of the two heterodyning radio frequency oscillators controlling pitch and volume. The electrical signals from the theremin are amplified and sent to the loud speaker. The same design principles were used by the inventor in the first proximity sensors which were successfully commercialized.

The theremin was envisaged by its inventor as a musical instrument for everyone. However, this dream has never materialized. The theremin went through ups and downs of its popularity, and it remains perhaps the most curious and fascinating musical instrument. It is also often considered as the most difficult to master. However, this is mostly due to a few problems which, as it was hypothesized in this research, can now be solved using computer technologies. Perhaps the main problem is luck of any tangible and visual contact with the instrument which requires the player to sense where exactly in the midair a particular note can be played. This is like singing where a perfect singer will not make any false note exactly knowing how to contract the vocal cords. The second problem is a non-linearity of the instrument, which has different distances between hand locations for different pitches: larger distances between lower pitches and smaller distances between high pitches. The third problem is that this is a monophonic instrument, and its sound can be only modified by changing its *waveform* and *brightness* (a.k.a. *color*). Last but not least, the theremins are very rare and expensive, while there are many other electronic musical instruments around invented during the past century. However still there are many people who fall in love with the magic of the music from the air, and the author of this paper is one of them.

Considering the theremin as the first hand-tracking device which was invented even before any computers, this paper investigates whether the theremin can be mimicked as close as possible using very different physical principles of optical hand-tracking and by adding further advantages of visual interaction. The ultimate goal of this research is to eventually fulfill the dream of the inventor to make the theremin a musical instrument for everyone and to prove that everyone can play music on it.

II. RELEVANT PROJECTS AND OPTICAL HAND TRACKING

There were very few attempts to simulate the theremin with a computer. For example, the relevant programming exercise presented in [1] was based on optical hand tracking

978-1-7281-2297-7/19/$31.00 ©2019 IEEE
DOI 10.1109/CW.2019.00018

and sound generation described in [2]. It produced different pitches following the position of one finger. There are also a few toy-like apps developed for iOS platform, e.g., "Therimax", which track a fingertip position on a touch-screen to convert it to pitch (vertical displacement) and volume (horizontal displacement) and web-based apps converting motion of a mouse cursor into audible sounds [3]. However, all these attempts hardly could be called successful due to various reasons. Perhaps the most important reason is that the developers did not have any experience of playing the real theremin and hence could not properly simulate it (while this paper author is an experienced theremin player). Usually, they try to squeeze into a limited tracking space as many octaves as possible to eventually play all the 88 standard piano pitches in twelve-tone equal temperament. There is also a challenge that to simulate sound in real time, the sampling frequency has to be at least doubled, and for the audible sounds (20 – 20,000 Hz) it becomes as high as 40,000 Hz. This sampling has to be done while concurrently tracking hands and computing the respective sound frequencies to eventually fill in a digital sound buffer to be played back.

There are a few affordable and precise optical tracking devices available, such as Leap Motion Controller, Microsoft Kinect for Windows (to be replaced with Azure Kinect in 2019), uSens Fingo, and BlasterX Sens3D. These devices cast light towards objects to determine their position which can be based on the principles of time-of-flight, structured-light, and stereo vision.

The time-of-flight cameras measure the distance based on the known speed of light and the time that takes the light to travel to the object and back to the sensor. Structured-light scanners measure the three-dimensional shape of an object using projected light patterns. The method of stereo-vision allows for tracking 3D position by analysing correspondence points in two images obtained by two cameras.

Structured-light and stereo vision data is then used to reconstruct the hand model based on appearance- and model-based methods. Appearance-based methods recognize gestures by matching the hand appearance in the image with a set of predefined gestures, however they suffer from the occlusion problem to precisely reconstruct positions of fingertips. In contrast, model-based methods allow for tracking of the whole hand rather than recognizing only certain gestures. Such methods work with 3D hand model by dynamically updating its parameters. These methods have achieved a good progress in recent years with an increased computing power and progress in machine learning which allowed to tackle the problem of finger occlusion [4, 5,6] and hand crossing [7].

Comparing the performance of the available optical tracking devices, Leap Motion controller may be called a winner in terms of a price-precision balance and relevance to the problem. With the price of USD 79.37 (Amazon US as of 30 May 2019), its dynamic accuracy is 2.5 mm according to [8] and the overall static and dynamic accuracy is below 0.5 mm according to [9]. However, the precision of tracking becomes affected when the fingers move close to each other. According to [10], it has an error of 0.878 cm when the distance between the thumb and index finger is 1 cm and the error becomes smaller than 0.04 cm when the separation distance is larger than 5 cm.

Though Leap Motion controller is mostly used in computer games, it was previously shown that it can be also used for precise geometric shape modeling tasks [11] by augmenting and even replacing common mouse-based interaction. It was therefore hypothesized in this research that if fine geometric shapes can be made by tracking hands in the air, the music can also be played in the air using Leap Motion controller.

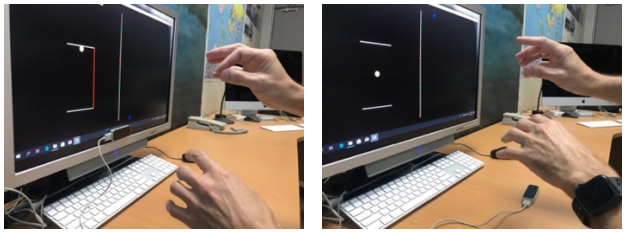
III. Challenges and Software Design Considerations

When playing the theremin, to avoid mutual interference the hands move in two different non-intersecting directions: dominating hand (pitch) – horizontally, the other hand (volume) – vertically. The pitch antenna senses the presence of moving objects within four feet, and the change of measured variable capacity is achieved by repositioning the hand, changing the finger configuration, and even by changing the position of the whole body. One common way of playing the theremin is to keep the thumb and the index finger of the dominating hand in a form of a ring while playing all the notes within an octave by extending other three fingers. Then, the lowest note in the octave is played when all the fingers are closed (Fig. 1b) while the highest note in the octave is played when the fingers are fully extended (Fig. 1c). Alternatively, the individual notes can be played by moving the whole hand from one position to another which, however, requires more precision from the theremenist. Even the whole body may participate in playing: by veering closer to the instrument the overall measured capacity changes, and it can be used to quickly jump one octave up or down. Otherwise the thereminists have to stand very still to avoid making false notes. For volume control, which is done by the other hand, usually displacement of the whole hand is used since the volume tracking distance is smaller than that for the pitch.

To mimic the theremin style of playing with Leap Motion, it has to be decided what exactly should to be tracked and what can be ignored. Leap Motion only tracks hands. It can track the position and orientation of the whole hand and/or any individual extended finger with a sub-millimeter precision. To achieve the optimal performance and to mimic the theremin style of playing as close as possible we track only five fingertips positions on each hand and then, as frequent as possible, compute two average distances from the fingertips of each hand to the device. This approach allows for using the theremin style of playing both with opening-closing fingers and/or with displacing the whole hand as well as any other possible styles. However, changing the body position can be used only if it changes the positioning of the

hands. Therefore, the player no longer has to stand still, and more expression can be shown with the body motion.

Both, horizontal and vertical placements of the tracking device can be used. In either case, one hand may move horizontally while another – vertically. However, the horizontal placement of the device appears to be more practical when working with the computer. In addition, with the horizontal placement of the device, both hands can also move vertically without creating a mutual interference (Fig. 2). This mode of playing also provides a possibility of placing the elbows on the desk to reduce fatigue, and it also allows for improving the precision by taking into account only vertical displacements of the fingertips and ignoring any horizontal displacements.



(a) (b)

Figure 2. Measured averaged distances from the fingers to the tracking device with (a) vertical placement and (b) horizontal placement are displayed by arrows.

Whatever way of tracking of the hands is used, the hand distances then have to be mapped to the pitch of the sound. In the real theremin, the pitch is continuous, as in the violin, rather than fixed, as in the piano. Since the theremin does not produce overtones, the beauty and fascination of its sound is achieved by subtle vibrations of the dominant hand so that the frequency is rapidly changing. This fast tracking of the hand can be produced with Leap Motion as well, however there may be a bottleneck within the tracking frequency limits which has to be at least 40,000Hz. Anyway, it was tested on several common desktop and notebook computers and worked well.

Additional advantages of using the computer are expected to be found in abilities to

- Make optional various playing styles;
- Provide mapping of hand distances not only to variable but to fixed pitches as well;
- Provide many digital filters for the simulated sound;
- Provide visual interfaces.

We anticipated that adding visual interfaces, which will display what pitch or note is being played, can make the digital simulation of the theremin much more user friendly, and will make it finally an instrument for everyone, as its inventor Leon Theremin always dreamed.

IV. Implementation Details

The digital simulation was implemented in C# using Leap Motion Software Development Kit (SDK) v.2.3.1 and an open source audio library NAudio by Mark Heath [12].

The position of the hands is tracked by using a dispatcher timer class that is triggered every 20 milliseconds. Triggering it faster than 20 milliseconds may not provide enough time for the simulation to generate a smooth sound. The Hand class, provided by the Leap Motion SDK, is used to get the position of the hands. Any dominant hand can be used for changing the pitch—it can be automatically detected. The pitch in MIDI standard is a number of a musical note from 1 to 88 with 12 tones in each octave. Mapping of the dominant hand position to the respective pitch value is done linearly, which already helps to play more accurately—in contrast to the real theremin the distances between pitches in the air when tracked by Leap Motion become the same. Unlike fixed 88 pitches of MIDI, the pitch value is computed as a real number which is then converted either to fixed or variable continuous frequency that has to be eventually sent to the wave generator for the sound output. The real value of *pitch* is converted to the sound *frequency* according to the twelve-tone equal temperament formula:

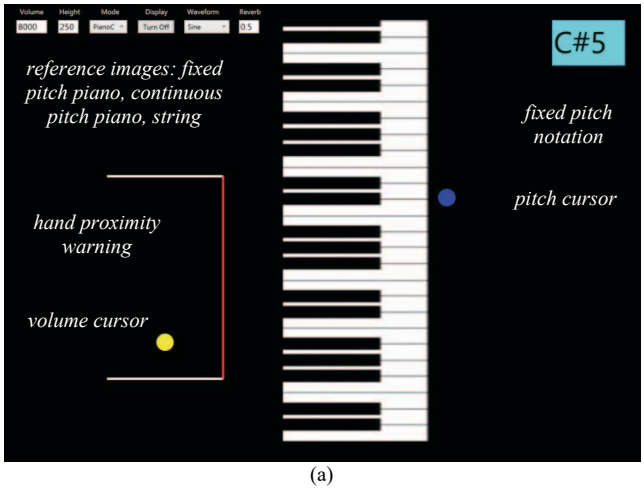
$$frequence = 440 \times 2^{\left(\frac{pitch-69}{12}\right)} Hz,$$

where 440 Hz is a frequency of A_4, which pitch number is 69. Thus, if variable continuous frequencies are required, as if when playing the real theremin, real values of *pitch* will be used. To play fixed notes as in the piano, integer *pitch* values have to be used. In a similar way, tracking distance of the other hand linearly maps to the value of the volume of the sound which can be an integer number. Various waveforms can be used to produce sound: sine, saw, square, triangle, etc. They are defined by mathematical functions. While filling in the sound buffer with frequencies based on one of these waveforms, both the frequencies and volume values have to be linearly interpolated to avoid "radio interference" effect which sounds as cracking noise. [2]. In addition, various digital sound effects, which also have to be mathematically defined to be implemented in real time, can be applied to the computed values such as reverberation, echo, low and high pass filtering, etc. After the buffer is eventually filled, it needs to be passed to an output device.

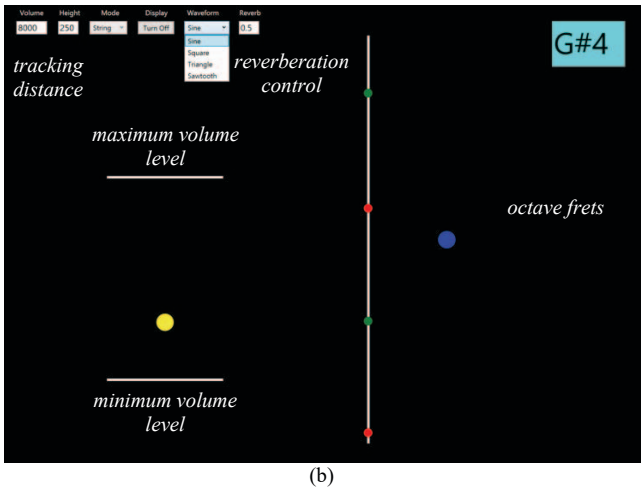
A few visual interfaces for the application were designed in such a way that they did not create a significant computational overhead for the hand tracking and sound generation tasks. The visual interfaces show the position of the user's hands with reference to one of a few visual displays of familiar musical instruments and the pitch of the sound the simulation is producing using their symbolic notations (C*n*, D*n*, E*n*, F*n*, G*n*, A*n*, B*n*).

For the players who are familiar with the piano, the piano keyboard can be displayed with a moving cursor indicating which pitch—fixed or variable continuous—is currently being played (Fig. 4a).

The range of the keys can be selected but typically 4 octaves are used and comfortably fits into 25-30 cm of the adjustable tracking distance ("Height" in the figure).



(a)



(b)

Figure 4. Configurations of one of the tested visual interfaces using as a visual reference (a) a piano keyboard and (b) a string with frets.

If the fixed pitches are being played, the cursor moves discretely pointing at the middle of the respective key (white or black). If the variable continuous pitches are being played, the cursor is merely used for indication in which area of the pitch range the sound is being played—it can help quickly move the hand i.e. one octave up or down. As an alternative, a vertical line representing a string with several frets is displayed to help playing fixed and continuous frequencies (Fig. 4b). The frets are located at the octave distance. In either interface selection, if the hands move close to each other, a visual warning (vertical red line) is displayed prompting to move hands farther from each other. Different waveforms (sine, saw, square, triangle), filters (reverberation, echo, low and high pass) and other settings can be also set, saved and restored, which allows for personalization of the interface for every player.

The developed application was offered for testing to ten users with and without musical background. It was then proved that with this rather elementary visual interface nine of them were able to play simple tunes within a very short period of learning in contrast to significantly larger time required for learning the real theremin or playing the digital theremin without the visual interface. It was also tested by the author who is an experienced theremin player. While he had to spent weeks before he could play anything descent on his first real theremin, it took just a few minutes to transfer this knowledge how play while looking at the cursors showing the hands positions.

V. Conclusion

We have proposed how to play music in the theremin style using Leap Motion optical hand tracking. We have also proposed visual interfaces which help make a quick start for the users, so that after just a short practice time they can continue playing with only occasional glancing at the monitor. This is still work in progress. We are now exploring how to further enrich playing music in the air by including polyphonic, multi-user and the whole body modes of playing as it was envisaged by Leon Theremin.

References

[1] D. Soshnikov, “Digital Termenvox based on Leap Motion” https://habr.com/company/microsoft/blog/214907, published in 2014 (15 May 2019).

[2] C. Petzold, "UI Frontiers - Sound Generation in WPF Applications", MSDN Magazine Blog, 25(2), https://msdn.microsoft.com/en-us/magazine/ee309883.aspx (15 May 2019).

[3] T. Szynalski, Online Tone Generator, https://www.szynalski.com/tone-generator/ (15 May 2019)

[4] X. Sun, Y. Wei, S. Liang, X. Tang and J. Sun, “Cascaded hand pose regression,” Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[5] L. Ge, H. Liang, J. Yuan and D. Thalmann, “Robust 3D Hand Pose Estimation in Single Depth Images Using Multi-view CNNs”: From Single-View CNN to Multi-View CNNs,” IEEE Transactions on Image Processing, 27(9): 4422-4436, 2018.

[6] J. Tompson, M. Stein, Y. Lecun and K. Perlin, “Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks,” ACM Trans. Graph., 33(5), #169, 2014.

[7] J. Taylor, et al., “Articulated distance fields for ultra-fast tracking of hands interacting,” ACM Trans. Graph., 36(6): #244, 2017.

[8] F. Weichert, D. Bachmann, B. Rudak and D. Fisseler, “Analysis of the Accuracy and Robustness of the Leap Motion Controller,” Sensors, 13(5): 6380-6390, 201.

[9] J. Guna, G. Jakus, M. Pogačnik, S. Tomažič and J. Sodnik, “An Analysis of the Precision and Reliability of the Leap Motion Sensor and Its Suitability for Static and Dynamic Tracking”. Sensors, 14(2): 3702-3722, 2014.

[10] R.L. Hornsey and P.B. Hibbard “Evaluation of the Accuracy of the Leap Motion Controller for Measurements of Grip Aperture,” Proc. of the 12th European Conference on Visual Media Production, ACM, #12, 2015.

[11] J. Cui and A. Sourin, "Mid-air Interaction with Optical Tracking for 3D Modeling," Computers & Graphics, Elsevier, 74: 1-11, 2018.

[12] M. Heath, “Audio and MIDI library for .NET”, https://github.com/naudio/NAudio (15 May 2019).