Development and application of eTheremin: hand tracking and AI technology facilitate educational and entertaining musical training

Tsen-Fang Lin
Dept. of Popular Music Industry
Southern Taiwan University of Science and Technology
Tainan, Taiwan
tflin@stust.edu.tw

Wen-Hsin Li
Dept. of Multimedia and Entertainment Science,
Southern Taiwan University of Science and Technology
Tainan, Taiwan
mb1k0117@stust.edu.tw

Abstract*—The rapid advancement of digital technology and artificial intelligence is driving significant transformations in the fields of music education and creation. This study addresses the challenges and hardware limitations of the traditional Theremin by introducing an innovative virtual Theremin system (eTheremin). The system integrates gesture tracking technology (MediaPipe) with an AI-powered timbre generation model (NSynth). It offers intuitive pitch and volume control while incorporating local soundscapes to promote cultural heritage and seamlessly combine education with entertainment, creating a novel musical interface. The research focuses on three primary objectives: 1. Exploring the educational impact of the virtual Theremin on pitch training. 2. Evaluating the developmental potential of gesture-controlled systems for musical performance. 3. Analyzing the integration of local soundscapes for cultural promotion and application value. This study is currently underway. Through experimental testing and user feedback, it examines the system's stability, learning effectiveness, and user experience. The findings aim to establish a new design framework, providing fresh insights for the broader application of virtual instruments in music education and creative practices.

***Keywords*—eTheremin, education, hand tracking, AI, soundscape**

I. Introduction

With the rapid development of digital technology and artificial intelligence, the landscape of music creation and education is undergoing significant transformation. Traditional methods of learning musical instruments are gradually being replaced by virtual and interactive technologies, providing music learners with more diverse opportunities. The Theremin, an instrument controlled through hand gestures to manipulate pitch and volume, has been widely admired since its invention for its unique contactless performance method and enduring popularity. However, the traditional Theremin's weight, cost, and the instability of self-assembled hardware present limitations for its broader application in music education.

This study leverages the Unity platform, integrating MediaPipe hand tracking technology and the NSynth artificial intelligence timbre generation model to design a virtual Theremin (eTheremin) with both performance and pitch-training functionalities. The system offers real-time gesture-based control of pitch and volume, supported by an intuitive operational logic—where the height of the right hand determines pitch and the vertical movement of the left hand adjusts volume. By aligning these features with music education theories and incorporating real-time keyboard position visualization, the system enables users, particularly beginners, to objectively and effortlessly develop pitch perception in practical scenarios. Furthermore, the system integrates soundscapes and cultural information from the Tainan region, combining music education with cultural promotion to create an innovative edutainment experience.

The objectives of this study include: (1) exploring the educational impact of the virtual Theremin on pitch training, (2) evaluating the musical performance potential of gesture-based controls, and (3) analyzing the practical value of integrating soundscapes with local cultural elements. This research aims to provide a novel design framework for the educational and creative applications of virtual instruments, promoting their broader adoption in the music domain.

II. Literature Riview

Since its invention by Léon Theremin in the early 20th century, the Theremin has garnered widespread attention in the fields of art, technology, and music due to its contactless performance technology, maintaining enduring popularity [1]. However, the hardware design of the traditional Theremin demands high sensitivity and stability of its antenna while typically being quite bulky and heavy, making it challenging to promote as a public teaching aid, particularly for small-scale educational purposes [2]. With the popularization of sensing technology, these limitations can now be addressed through virtualization technology, enabling broader accessibility. For example, smartphones or regular cameras can be used to capture hand gestures, with tracking systems simulating the playing of the Theremin. However, most existing research has yet to integrate a wider variety of timbre options or fully utilize the Theremin's unique educational potential.

Virtualization technology aids Theremin practice by reducing errors and making the learning process more intuitive [3]. Additionally, correlating hand height with pitch—where higher hand positions correspond to higher pitches—aligns with the well-known logic of pitch perception in Kodaly's music education theory. Furthermore, integrating visual keyboard displays can enhance learners' spatial memory and immediate feedback capabilities for pitch locations, further improving pitch training outcomes. Artificial intelligence technologies have also been widely applied in music generation and analysis, with the NSynth model's advantages in multi-timbre generation being well-established [4]. Research suggests that AI-generated musical

content can dynamically and engagingly attract learners, allowing them to develop musical perception skills effortlessly [5]. However, current research primarily focuses on timbre-related aspects and has not fully explored practical applications that integrate humanistic dimensions, such as using AI-generated audio with the Theremin for pitch training tools or combining soundscapes with local cultural elements.

Soundscapes refer to recordings of natural sounds and human environments from specific regions. These sound materials are rich in cultural significance, deepening the contextual meaning of music education. Studies indicate that applying soundscapes in music creation and education not only helps learners understand the relationship between sound and the environment but also fosters cultural identity and supports cultural preservation [6]. In recent years, soundscape-related topics have gained increasing attention in Taiwan, with many cases successfully incorporating soundscapes into music teaching to achieve positive outcomes. However, these studies have not yet integrated artificial intelligence technology to further enhance their applications.

Soundscapes refer to recorded natural sounds and human environments from specific regions, forming sound materials rich in cultural significance that can deepen the contextual meaning of music education. Studies indicate that the application of soundscapes in music creation and education not only helps learners understand the relationship between sound and the environment but also fosters cultural identity and supports cultural preservation [6]. In recent years, topics related to soundscape have arisen in Taiwan, and there are many cases where they have been integrated into music teaching and achieved good results. However, these studies did not integrate artificial intelligence technology to further enhance its applications.

Multisensory learning theory suggests that synchronized visual, auditory, and kinesthetic stimuli can produce stronger and more efficient learning outcomes [7]. Virtual instrument designs that combine gesture control, visualized keyboards, and dynamic auditory feedback provide multidimensional support for music education. Based on this theory, the eTheremin designed in this study emphasizes intuitive gesture control of pitch and volume while integrating soundscapes to seamlessly merge cultural and educational objectives.

III. Methodology

The system aims to enhance users' sound perception through virtual eTheremin game-like somatosensory methods, including pitch perception and instrument recognition, as well as plans to connect artificial intelligence sound libraries and soundscape adjustments (Fig. 1-4). The following is the structure of system integration, including parts A-E below.

a) Hardware and Software Environment

Platform and Tools: The system is developed using Unity 2021.3.0f1, integrated with the MediaPipe Unity Plugin for gesture tracking. It runs on a Windows 11 environment, leveraging an RTX 4070 GPU to support real-time computation and sound generation. Audio data is sourced from the NSynth model and stored in a local sound effects database.

Fig. 1. Example of system operation eTheremin performance

b) Virtual Theremin Modules

Performance Mode: The right hand controls pitch, while the left hand controls volume. Users can switch timbres via buttons or ropdown menus.

Pitch Training Mode: Players perform gestures to match the correct pitch and timbre as prompted by the system.

Soundscape Functionality: Integrate the soundscape of Tainan's scenic spots and display cultural and geographical information to enhance user engagement. In addition, local flavor icons and photos of the location where the soundscape was recorded are displayed to enhance the user's local cultural experience (Fig. 5).

c) Sound Data Processing

Sound Generation: The NSynth model is used to generate eight major timbre categories (e.g., guitar, keyboard, synthesizer) and their subcategories. The sound effects are encoded into JSON format, mapping pitch, volume, and timbre information.

Dynamic Control Logic: Unity scripts are implemented to respond in real time to gesture data transmitted by the MediaPipe Plugin, dynamically adjusting pitch, volume, and timbre.

Fig. 2. eTheremin user interface diagram

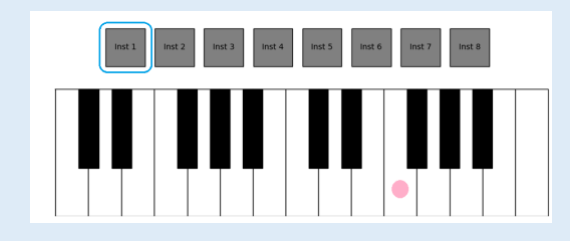

Fig. 3. Users can see their own performance

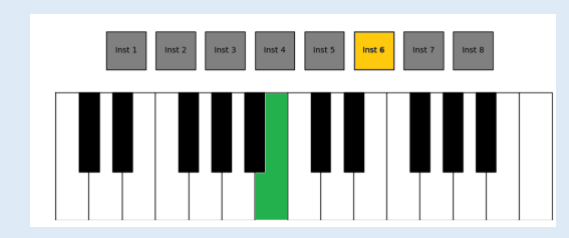

Fig. 4. The correct answer is presented at the end of the game

Fig. 5. Example of soundscape display in user interface with local characteristics

d) AI Applications

Player Behavior Data Collection: Tracks players' accuracy and reaction time during pitch training mode, utilizing AI to analyze individual performance and adjust game difficulty accordingly.

Personalized Learning Path Design: Automatically generates tailored pitch training content based on AI model results to meet individual player needs.

e) Experimental Design

Participants: Divided into preliminary expert test reports and general user test results. The current stage is expert testing. 10 experts with music background have been tested. It is expected that a total of 30 professionals with music learning background will participate. Statistical data analysis will be conducted, and the system will be modified based on opinions. After that, 30-50 players with non-music background will be opened for testing, which is the goal of the second phase (Table I, II).

Test Items:

- System stability and sound response speed.
- Evaluation of learning outcomes in the pitch training mode.
- Effectiveness of the soundscape functionality in promoting cultural education.

Data Collection and Analysis: A combination of qualitative and quantitative analysis is conducted through questionnaires and system-generated records.

IV. Results and Discussion

1. System Performance Evaluation

The current latency for pitch and volume control falls within an acceptable range for users' natural perception. Further rigorous objective evaluations will be conducted to reduce the average control latency during performance to less than 30ms, meeting the requirements for real-time performance. Additionally, efforts will be made to stabilize the timbre-switching speed to within 50ms to ensure a seamless user experience without noticeable lag.

TABLE I. General User Questionnaire Form Design (Currently)

	Question
A	Have you had any prior experience with musical instruments? (Yes/No)
A	How familiar are you with digital or virtual musical tools? (Not familiar - Very familiar, 1 – 5 scale)
B	How easy is it to control pitch using hand gestures? (1 – 5 scale: Difficult - Very easy)
B	How easy is it to adjust volume using hand gestures? (1 – 5 scale: Difficult - Very easy)
B	Do you find the interface intuitive and easy to use? (Yes/No)
C	Did the pitch-training mode help you understand pitch perception better? (1 – 5 scale: Not at all - Very much)
C	How engaging is the soundscape feature that incorporates Tainan's cultural elements? (1 – 5 scale: Not engaging - Very engaging)
C	Would you recommend this system to someone interested in learning music? (Yes/No)
D	What did you enjoy most about the system? (Open-ended)
D	What challenges did you encounter while using the system? (Open-ended)
D	What improvements would you suggest? (Open-ended)

(A: Background Information, B: Usability, C: Learning and Engagement, D: Feedback.)

TABLE II. Expert Questionnaire Form Design

	Question
A	What is your primary professional focus? (e.g., music education, music technology, performing arts, other: specify)
A	How familiar are you with the traditional Theremin? (Not at all familiar - Very familiar, 1 – 5 scale)
B	How do you rate the accuracy of pitch control in the virtual Theremin compared to traditional instruments? (1 – 5 scale: Poor - Excellent)
B	How satisfied are you with the timbre-switching functionality in terms of speed and variety? (1 – 5 scale: Not satisfied - Very satisfied)
B	Do you find the integration of Tainan's soundscape and cultural elements enhances the educational value? (Yes/No)
B	Are the gesture controls intuitive for performance and training? (1 – 5 scale: Not intuitive - Very intuitive)
C	Do you believe this system has potential applications in music education? (Yes/No)
C	What features would you recommend to improve its application in music education? (Open-ended)
C	How do you assess the system' s potential for creative performance use in live settings? (1 – 5 scale: Poor - Excellent)
D	Does the AI-assisted personalized learning path address individual user needs effectively? (1 – 5 scale: Poor - Excellent)
D	How useful is the real-time feedback in guiding users during pitch-training sessions? (1 – 5 scale: Not useful - Very useful)
D	Do you find the soundscape-based features engaging for cultural learning? (Yes/No)

(A: Background Information, B: System Functionality, C: Educational and Creative Potential, D: AI Integration.)

2. User Experience Feedback

Multiple participants with a musical background noted that the virtual Theremin offers superior pitch control accuracy compared to its physical counterpart. They also highly appreciated the diversity of timbre options.

Planned collaboration with faculty from the General Education Center will enable broader testing among non-musical background users. Feedback and results will be collected through questionnaires, providing insights into general user experiences. It is anticipated that the pitch training mode will enhance users' musical perception

abilities and receive recognition for its edutainment design.

3. Effectiveness of AI Applications

After public testing, the system will analyze user experiences to evaluate the percentage improvement in pitch identification accuracy and timbre selection accuracy facilitated by AI-assisted edutainment features. The differences between the two will be observed, and the impact of personalized learning paths on improving user interest and focus in music training will be assessed. Questionnaire feedback will include users' impressions of the integrated audiovisual interaction experience provided by the soundscape functionality. If predominantly positive responses are received, the system will be deemed effective in subtly fostering users' understanding and recognition of Tainan's local culture.

4. Research Contributions and Future Outlook

This study integrates AI and gesture control technologies, pioneering new applications for the virtual Theremin. The system not only serves music education purposes but also shows potential for expansion into performing arts and cultural promotion. Future plans include incorporating more soundscape data and exploring the use of deep learning models to further enhance the intelligence of pitch training.

V. CONCLUSION

This research addresses these challenges by designing and implementing a virtual theremin system (eTheremin) based on gesture control and artificial intelligence technology that can overcome the disadvantages of pitch control and hardware stability of traditional theremin. From this point of view, the system's intuitive operating logic, real-time keyboard prompt show practical activities and timbres offered in a mechanical and sharing way make novices have the opportunity to develop the sense of pitch in immersive activities. Additionally, the system integrates aspects of Tainan's local soundscape and culture to serve as an effective prop that not only promotes music education but also helps to enhance cultural awareness and identity. Preliminary test results show that users with professional music backgrounds gave good evaluations of the system's pitch control accuracy and timbre diversity, and users with non-music backgrounds also expressed enjoyment in the game-based learning mode. The AI-driven personalized learning path design further enhances users' learning engagement and music perception. In the future, the system will continue to optimize response speed and timbre switching performance and combine deep learning technology to further improve the level of intelligent pitch training. In summary, this study opens up new possibilities for the application of virtual instruments in education, creation, and cultural promotion, and provides a reference model for subsequent research and development in related fields.

REFERENCES

[1] P. Nikitin, A. Parks, and J. Smith, "RFID-Vox: A Tribute to Leon Theremin," 10.1007/978-1-4419-6166-2_16, 2013.

[2] T. -F. Lin and C. -H. Yang, "Intuitive, Interactive Training System for the Sense of Sound Using Portable Theremin Device," *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Osaka, Japan, 2024, pp. 108-111, doi: 10.1109/ICAIIC60209.2024.10463303.

[3] David J., D. Damian, and G. Tzanetakis, "Evaluating the effectiveness of mixed reality music instrument learning with the theremin," *Virtual Reality*. 24. 10.1007/s10055-019-00388-8, 2020.

[4] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D. & Simonyan, K., "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," *Proceedings of the 34th International Conference on Machine Learning*, in *Proceedings of Machine Learning Researc*h 70 : 1068-1077, 2017.

[5] Dai, Shuqi. "Towards Artificial Musicians: Modeling Style for Music Composition, Performance, and Synthesis via Machine Learning." Diss. Stanford University, 2023.

[6] Aktaş, Rodem. "Environmental Soundscapes: Soundmapping in Music Education," 2024.

[7] Rao, A. Ravishankar. "An oscillatory neural network model that demonstrates the benefits of multisensory learning." *Cognitive neurodynamics* 12(5), 2018: 481-499.