Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220645

Artificial Intelligence-based Musical Instrument Accompaniment System

Kevin P Chen

Shanghai American School, Shanghai 201106, China

kevin02px2023@saschina.org

Abstract. The string learning process requires students to be able to accurately grasp the accuracy of notes and rhythms in the process of playing, and its evaluation is usually done by instructors, but long one-on-one instruction is difficult to achieve in actual teaching. In this project, through the research of audio hardware systems and digital signal processing software technology, I design a string performance recognition robot system combining hardware and software, to realize automatic accompanying practice. The hardware of this system consists of a Raspberry Pi card-type computer, recording equipment, and cueing equipment. The software system processes the audio signal collected by the microphone through artificial intelligence technology and digital signal processing technology to realize the recognition of notes, intensity, and other elements according to the frequency characteristics of string instruments. The accuracy of the user's performance is judged by comparing the recognition result with the content of the score, and the user is guided by real-time hints through the display system. The experimental results show that the system can recognize 99% of the string audio with good recognition stability, so it can give accurate and timely performance guidance to the player.

Keywords: Robot, Companion, Intonation, Rhythm.

1. Introduction

Musical instruments are a hobby for many people, and learning an instrument is also a need for many people. In the process of learning a musical instrument, learners want to always have a teacher to help them at all times. Therefore, I want to design a device that can help students to practice cello. I name this device “Alpha DoMi”. This device can record the mistakes made by students during practice, including pitch, rhythm, volume, and so on [1]. When a student makes no mistakes in practicing, the system will reward him/her accordingly; when a practitioner makes a mistake in playing, the device will promptly give a corresponding prompt. This system can help the practitioners to improve their practice efficiency effectively [2].

I tried to build such a system based on deep learning technology to achieve the goal of helping the practitioners to identify the pitch [3]. The system needs to have the following four functions:

A. Recognize the music score and determine the current music piece played;
B. Obtain audio information and identify the accuracy of the performance in real-time;
C. Judge the accuracy of the playing action based on the video information during the playing;
D. Give suggestions for improvement based on the performance.

Following the above requirements for the system, this paper will firstly introduce the design of the instrument accompaniment system; secondly, this paper will conduct an empirical study to check the

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

effectiveness of the system; finally, I will summarize the shortcomings of the existing system and give the corresponding solutions.

2. System Design

The design of the musical instrument accompanying system includes the selection of the microcontroller, the selection of the sensor, the design of the detection system, and the appearance design.

2.1. Microcontroller

The Raspberry Pi 4 is the only microcontroller that meets the requirements of this design, while the Arduino, Raspberry Pi 4, and micro-bit are unable to meet the requirements. Compared to Raspberry Pi 3, the Raspberry Pi 4 has a much higher processing speed, can output video at 60Hz, supports dual monitors, uses a Cortex A72 CPU core, 4GB of RAM, and the processor can guarantee the speed of audio recognition.

2.2. Sensor

In terms of sensors, there are two main types of sensors, one for audio recognition and one for image acquisition.

For the image sensor, the system needed a high-definition camera to scan the music score, instruments, and playing behavior. I chose the original Raspberry Pi camera, which has 5 megapixels, and tested it to be able to perform the image acquisition tasks the system needed.

For the audio acquisition sensor, the system needed to be able to accurately identify the sound made by the player playing the cello. Finally, the system uses a WM8960 module with two high-quality silicon microphones on board, which can record left and right channels and be controlled by Python.

2.3. Detection System

In terms of the recognition system, 3 major functions need to be achieved.

The first is the recognition of the score, which is mainly a text recognition system, i.e. an OCR system. This system needs to identify the text information of the score from the page of the score and retrieve the score object in the database by this text information, to realize the alignment of the score information [4].

Secondly, it is about the recognition of the performance audio. After I obtain the audio information of the performance scene, I need to operate filtering the current performance information to achieve the alignment of the scene research situation with the audio information in the database [5]. I constructed the corresponding models using the BERT model as well as the partial time series model. After the operation, I compare the audio information of the score with the audio information in the database and identify the part that deviates a lot [6,7]. This part is usually the part that needs to be improved.

Finally, there is the system for action recognition. I used a skeleton node recognition scheme. After finding the parts with large audio deviations, I proceeded to skeletal node movement analysis for this part of the movement, and through this analysis, I tried to find the movements where the player played incorrectly.

2.4. Appearance Design

The product design includes sketch design, modeling design, and finished product.

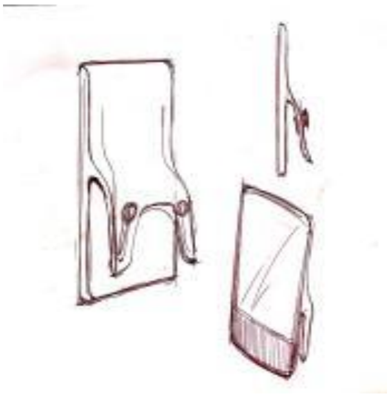

Figure. 1. Sketch design 1. This was designed to be hung on the music stand while having screws behind it to stabilize it.

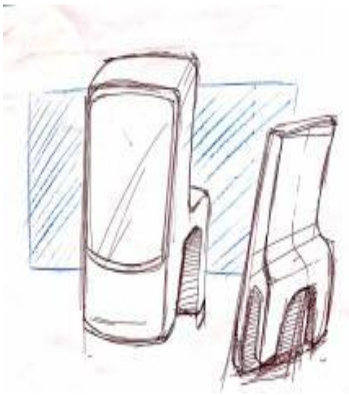

Figure. 2. Sketch design 2. This design increased the length of the back part to better support the robot on the stand.

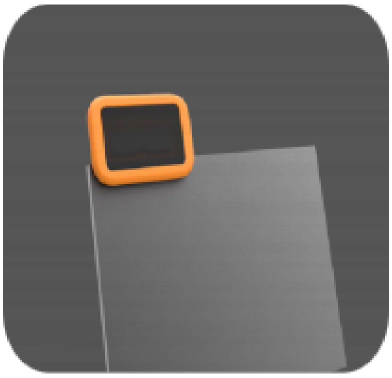

Figure. 3. Modeling design 1. This design took into consideration of the screen size and the size of the robot itself so that it wouldn’t take up the entire stand.

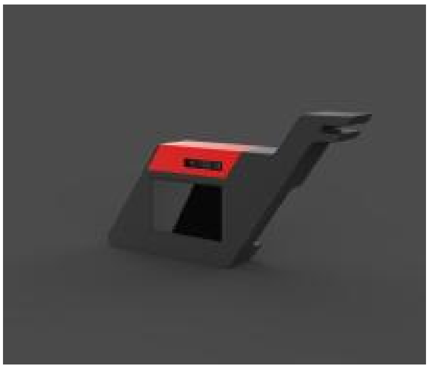

Figure. 4. Modeling design 2. This design attempted to add some decorations to make it look more artistically pleasing.

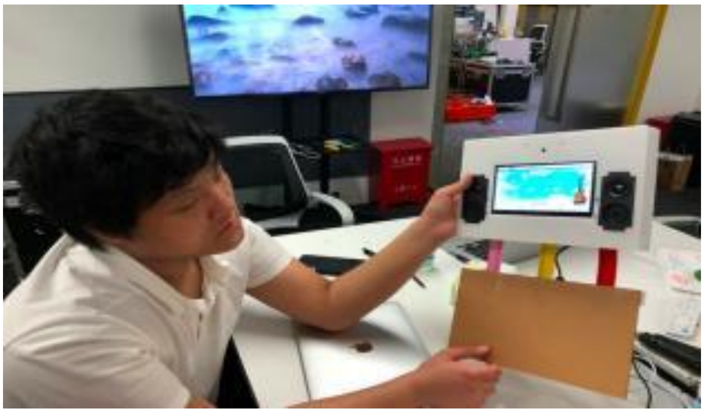

Figure. 5. The macro button chooses the correct format automatically.Finished product. This design was the final product that is being presented.

3. Sound Measurement Experiment

With the development of DSP noise processing technology, the use of this technology can improve the recognition of cello music, decompose the cello music signal, and combine the fuzzy information sampling method for cello music signal output conversion control, to achieve the optimal recognition of cello music signal, thus improving the cello training effect, and the practical application effect of the method is verified through experiments. The cello music signal recognition based on DSP noise processing, firstly, needs to construct a cello music signal spectral feature sampling model, and combine the multi-dimensional information feature decomposition method to decompose and detect the features of the cello music signal, use DSP noise processing cello music signal, in the channel model, carry out the cello music signal output stability control, use fuzzy modulation method, carry out synchronous demodulation control of cello music signal, and improve the synchronous conversion control of cello music signal.

In the experiment, the acoustic signals of the A and D strings of the cello were collected separately, and the correct posture and two common wrong postures were used in the performance. The two incorrect postures were typical and common mistakes made by the students during the practice of air strings. The specific postures are:

Posture 1: Playing with an empty string, using an incorrect string position, where the touchpoint between the bow and the string is too high, the bow does not fit the string, the bow pressure on the string is not achieved, and the bow speed is difficult to control;

Posture 2: Playing with an empty string, using an incorrect bow grip, where the thumb is stuck inside the bow stick and touches the fingers holding the bow stick in front, extreme tension, deformation of the hand shape, extreme difficulty in bow change, excessive bow to string fit, excessive bow pressure, too fast bow speed.

Posture 3: Playing with an empty string, holding the bow, and touching the string in a reasonable position, where the touching point is close to the gauge, the bow is held loosely in the right hand, and the bow is fully attached to the string, with normal bow pressure and bow speed.

The results I obtained after processing the raw audio in three different ways.

Table 1. The results after the original audio were identified.

Algorithm 1	Original audio comparison experiment (50 notes)				
Scenes	Experiment 1 Exact number	Experiment 2 Exact number	Experiment 3 Exact number	Experiment 4 Exact number	Mean
Silent	46	49	45	40	.9
Classroom	40	42	41	38	.805
Theatres	42	41	39	37	.795
Public	36	24	32	40	.66

Table 2. Results of the original audio processed by the "string frequency feature recognition algorithm" and then recognized.

Algorithm 1	String frequency feature recognition algorithm (50 notes)				
Scenes	Experiment 1 Exact number	Experiment 2 Exact number	Experiment 3 Exact number	Experiment 4 Exact number	Mean
Silent	47	48	50	49	.97
Classroom	47	41	44	42	.865
Theatres	42	46	39	45	.86
Public	36	32	33	34	.675

Table 3. Results were obtained by combining the string frequency feature recognition algorithm and the DSP algorithm and then performing the recognition.

Algorithm 1	DSP noise filtered string frequency feature recognition algorithm (50 notes)				
Scenes	Experiment 1 Exact number	Experiment 2 Exact number	Experiment 3 Exact number	Experiment 4 Exact number	Mean
Silent	50	49	49	50	.99
Classroom	50	46	47	48	.95
Theatres	46	48	47	48	.945
Public	38	36	37	35	.73

4. Teaching experiments and conclusions

The method was applied to the teaching of the cello by using an artificial intelligence-based instrument accompaniment system to collect, analyze, and visualize the signals of cello sounds in different playing postures, and then to obtain the above relationships. A comparison experiment was conducted with 10 beginners in each lesson: (1) 5 of them were taught with conventional teaching methods, and 3 of them were able to maintain the correct posture and produce the basic sound satisfactorily, while 2 of them were unable to maintain the correct posture and produce the basic sound satisfactorily. (2) Another 5 students used the artificial intelligence-based instrument accompanying system to analyze and compare the sound produced by the correct method with the sound produced by the wrong method. The students can quickly find the basic correct posture by observing, listening, and feeling. In the actual performance, the percentage of students who maintained the correct posture and produced the basic sound satisfactorily was 80%, which was 20% higher than the original oral instruction.

After analysis, the main reason for the improvement of the teaching effect is that by using the artificial intelligence-based instrument accompanying system, students can use the three ways of hearing, body feeling, and observing the spectrum display to give feedback on the correctness of playing posture, so that they can correct and adjust their playing posture in time, improve their playing effect and produce a basic satisfactory sound. In contrast, the traditional teaching method can only use the first two types of feedback, so it is not easy for students to understand the content of the teacher's explanation of correct posture, and the ability of different students to understand the difference between them, which will lead to uneven learning results.

5. Improvement and Prospect

The AI-based musical instrument accompaniment system has been able to meet everyone's needs for accompaniment when practicing cello, but some improvements can be made.

A. At the same time, the experimental results also show that the recognition rate in a noisy public environment is difficult to improve significantly, which can be achieved in the future through software algorithms and hardware equipment enhancement;

B. Develop a voice feedback function, i.e., use voice to prompt the piano practitioner;

C. Set up a scale of intensity from 1- 10, ppp, pp,p,mp,mf, f,ff,fff, to enable the robot to automatically recognize and distinguish the student's playing. If possible, the intensity f fading to p can be achieved, from f dim. to p;

D. Automatic rhythm adjustment, in a piece, the rhythm is not a rhythm from beginning to end, such as 80- 100- 140, so that the artificial intelligence-based instrument accompaniment system can recognize the rhythm;

E. Musical expression terms: such as Allegro in fast plate, Moderato in medium plate, Andante inline plate, Largo in a wide plate, etc., there is a range of speed in music, and after the robot scans the score, the robot should be allowed to understand the meaning of different playing speeds to reflect that the robot is different from the metronome;

F. Identify legato, break and skip by programming;

G. Give an example of a piece played by a master as a sample, and use it as a scoring criterion. The product design includes sketch design, modeling design, and finished product.

6. Conclusion

In short, this robot will be able to competently accompany the performer while he/she is playing at 99% accuracy with careful selection of the parts and the use of DSP.

References

[1] Huang, Allen, and Raymond Wu. "Deep learning for music." arXiv preprint arXiv:1606.04930 (2016).

[2] Briot, Jean-Pierre, Gaëtan Hadjeres, and François-David Pachet. "Deep learning techniques for music generation--a survey." arXiv preprint arXiv:1709.01620 (2017).

[3] Kereliuk, Corey, Bob L. Sturm, and Jan Larsen. "Deep learning and music adversaries." IEEE Transactions on Multimedia 17.11 (2015): 2059-2071.

[4] Sturm, Bob L., et al. "Music transcription modeling and composition using deep learning." arXiv preprint arXiv:1604.08723 (2016).

[5] Haralick, Robert M., et al. "Pose estimation from corresponding point data." IEEE Transactions on Systems, Man, and Cybernetics 19.6 (1989): 1426- 1446.

[6] Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[7] Murphy-Chutorian, Erik, and Mohan Manubhai Trivedi. "Head pose computer vision: A survey." IEEE transactions on pattern analysis intelligence 31.4 (2008): 607-626.