www.cadanda.com
Computer-Aided Design AND APPLICATIONS

Gesture Interactive Music Playing System based on Internet of Things and its Realization Method

Jun Zhao[1*] and Marianne Zhao[2]

[1]Department of Music, Zhengzhou Preschool Education College, Zhengzhou, Henan 450099, China, drzhaojun@hotmail.com
[2]Department of Music, Zhengzhou Preschool Education College, Zhengzhou, Henan 450099, China, mlaumets@gmail.com

Corresponding author: Jun Zhao, drzhaojun@hotmail.com

Abstract. This paper discusses the implementation and design of a music system according to the Internet of Things. The system uses ARM920 TS3C2410 embedded processor and WinCE OS to realize the central control module, and then builds a sensor network according to the ZigBee2007/PRO network standard, and reads information from sensors such as light, sound, infrared sensors, and temperature through RFID radio frequency identification technology Take, and finally combine the principles of music psychology and user behavior log to design the music selection logic, and finally select the best music for the user to improve the user's mood, improve their quality of life and work and study efficiency. At the same time, the system uses voice recognition technology to enhance user interaction.

Keywords: Internet of Things; Gesture Interaction; Music Play System
DOI: https://doi.org/10.14733/cadaps.2022.S6.58-67

1 INTRODUCTION

The computing model changes every 15 years. This view, known as the "fifteen-year cycle law," was considered as accurate as Moore's Law once it was put forward by the former CEO of IBM Gerstner. Throughout history, the changes occurred around 1965. of personal computers in 1980, the popularization were major changes in the history of computer development. Kang and Seo [1] think the "Internet of Things" was born in 2010, it has attracted much attention and has become a hot topic in the industry. Gómez et al. [2] think that it is widely called the third wave after the information industry has relayed computers and the Internet.

As more and more sensor devices are connected to the Internet, there is access to the Internet all the time in every field. These behaviors generate a large amount of real-time data, so it is called the new concept of Internet of Things (IoT) Came into being. As early as 1999, Kevin Ashton

proposed the initial concept of the Internet of Things. He envisioned that all electronic devices in the world will be connected to the Internet in the future, and they will have their own electronic tags. This will extend the current Internet to the ubiquitous areas around us. Kind of objects. In 2005, at the World Summit on the Information Society held in Tunisia, the International Telecommunication Union issued a report formally establishing the concept of the Internet of Things, defining it as: "ubiquitous network", "dynamic network", and "ubiquitous computing". At the time when communication technology is becoming more and more mature, people's lives are surrounded by sensors and mobile devices. The development of the Internet of Things is closely integrated with daily life, bringing ubiquitous impacts to people's lives. The Internet of Things technology has been applied in many fields, such as autonomous driving, digital home and so on. At present, there is a typical paradigm Trigger-Action Pattern (TAP) for IoT data applications. TAP can realize cross-domain item interaction. Users can edit Recipes according to their own needs and combine functions of different items. The application research of Recipes focuses on designing a new framework to make the creation and use of Recipes more convenient and intelligent. At the same time, the success of IFTTT and Inte-gromat proves the usability of the TAP mode. With the rapid development of the Internet of Things, the number and interaction of sensors and smart devices in the network are constantly increasing. Daily life is covered by various complex and diverse Internet of Things data. All kinds of Internet of Things data have their own unique attributes, and within a certain range of Internet of Things data form a complex system that blends and connects with each other. Faced with such an intricate system, it is often difficult for users to quickly and accurately find the resources they need. Therefore, it is necessary to study efficient and intelligent methods to mine the information among IoT data, and data mining provides researchers with corresponding methods.

At present, the research on the data processing of the Internet of Things is increasing. Due to the heterogeneity of the Internet of Things data and the interoperability of items, the development of application programs and the processing of data have caused difficulties. Therefore, related technologies such as linked data and ontology have been applied. Solve the heterogeneity and interoperability of IoT data. Ganzha M et al. realized the interoperability between multi-domain IoT platforms and explained which ontology can be used for interoperability in the development of cross-domain IoT platforms. Chen et al. proposed an item recommendation model, based on user preferences, mapping the social relationships of items in a low-dimensional space, generating the social similarity of items, and using writing filtering to generate a recommendation list. Noura et al. designed a semantic framework that can automatically extract key subject information from relevant documents in the Internet of Things application field. In recent years, the new paradigm trigger execution mode (Trigger-Action Pattern, TAP) has been widely used in the actual scenarios of the Internet of Things. In this mode, users can edit TAP rules (often called Recipes) according to their own needs, and use corresponding functions Combine and link items. The current research on TAP is mainly divided into the syntax analysis of Recipes and the application research of Recipes. Liu et al. designed an end-to-end neural network that effectively utilizes natural language structures by calculating the weights of words in two stages, and automatically translates Recipes into programs. In order to make the creation of Recipes faster, the paper gives a novel user interface to select user ingredients to help users create Recipes. Corno et al. annotated the words in Recipes and used Semantic Web technology to improve the expressive ability of Recipes. It can be seen that the previous research focuses on the natural language processing of the recipes in TAP, but there are few related researches on the use of heterogeneous information network models to mine TAP containing multiple types of structural information. This article gives a TAP The clustering method of heterogeneous Internet of Things data uses the analysis of the heterogeneous information network constructed by TAP rules to mine the potential structural information in the Internet of Things data.

Baris et al. [3] think the basic function of a music system according to the Internet of Things is to collect information such as light, sound, infrared, and temperature in the environment through various sensors, and upload these sensor information to the central control system, which analyzes and processes the information, and Combined with the research results of music psychology, the appropriate music is selected and transmitted to the remote wireless speaker via wifi, so as to play

the background music most suitable for the user's mind, and improving work efficiency. At the same time, Keivanpour and Kadi [4] think users can also share their own music data, comment on songs, view song information, etc. on the website provided by this system.

Lee et al. [5] think existing voice interaction systems, such as smart phones, smart speakers, smart watches, etc., have the following steps in their use process: Step one, wake up the system, tap or slide on the touch screen to enter the voice assistant interface, or voice speaking When the wake-up language is heard, the system will enter the voice assistant interface after hearing it; step two, wait for feedback, at this time the system with a display screen will display the voice assistant interface on the screen, and the system without a display screen will give voice feedback to inform the user of the voice assistant The mode has been activated; step three is to send a voice command. The user must complete the first two steps before issuing the actual voice command. Because voice wake-up and voice feedback require a certain amount of time, the consequence of this interactive method is that every time the user sends a voice command, no matter how long or short, it must go through a waiting process of several seconds, and there can be no voice commands in between. Longer pause, otherwise you need to go through step one and step two again. Such voice interaction methods cannot meet people's requirements in the following scenarios: In some emergency situations, users need to use voice to quickly convey emergency information such as distress to others; users need to use the voice function through the system multiple times in a short period of time; in some emergency situations In a noisy environment, the success rate of the voice wake-up system is not high. Therefore, a faster and more direct way of voice interaction is needed [6,7].

The 21st century is the era of information technology. The development of various communication technologies and automation control technologies has promoted the progress of human civilization, and the living environment has gradually become more comfortable and safer. In recent years, according to statistics released by IDG, smart home appliances will increase by 5-10 times in the future. At present, there are various types of smart home control products on the market, with different implementation methods, but they are limited to high-end consumption. Accepted by most families, and there is no uniform standard for interface control. It is the common pursuit of all developers and users to develop a low-cost, simple and easy-to-operate remote monitoring system. The home appliance manufacturing industry at home and abroad is in a period of transition from traditional home appliances to smart home appliances. The application of communication technology and automatic control technology to home appliances, and the development of smart home appliance systems are currently a hot research topic [8-10].

2 SMART MUSIC SYSTEM MODULE DESIGN

Through a certain algorithm, select the song that composites the scene at the time. The structure diagram of the system is shown as in Figure 1. The main function is to install a pair of infrared sensors on both sides of the door, which will collect whether someone enters the room, which performs further processing. The infrared sensor is the ht-m series of beam sensors. The realization principle is that the voltage difference between the two ends of the door is maintained at 5V under normal conditions. If a person passes through the door, the infrared ray emitted by the sensor is blocked, and the electric music at both ends is also blocked. Back to the change, the information needs to be transmitted through Wi-Fi at this time. At this time, the data collector installed on the computer will convert the analog signal into a digital signal according to a pre-defined protocol (for example, in this system, 0 when no one enters, and 1 when someone enters), convert the analog signal into a digital signal, and send it to the central office. Control the system to complete the collection of infrared signals.

It is simple. In other words, the photosensitive sensor uses the principle that the resistance of the photosensitive resistor changes due to the influence of the light intensity to send an analog signal of the light intensity to the robot host. We also use the data collector to transmit the information from the light sensor according to a predefined protocol (for example, in this system, the voltage value is divided into 5 levels, which are represented by numbers 1-5, and then different

light intensity values are shown.) To the central control system, which processes the information to complete the collection of light perception information. For example, the system can recognize "next song", "previous song", etc., and can perform corresponding processing.

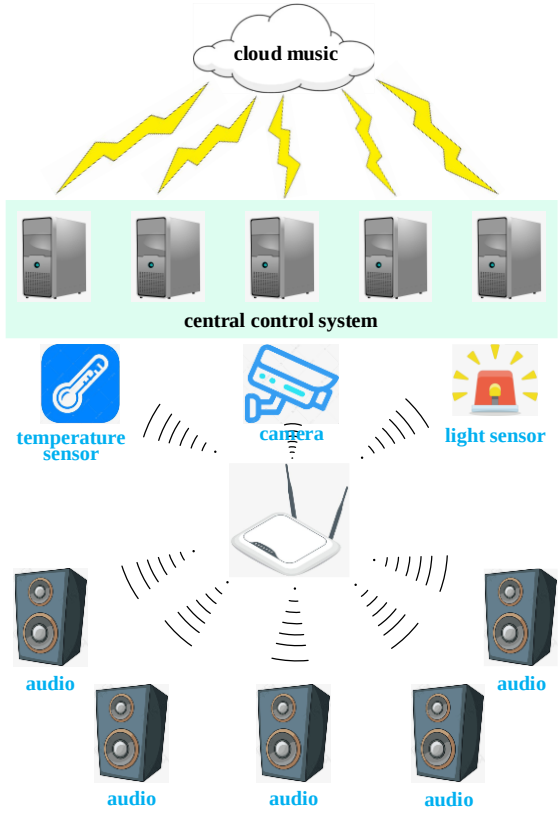

Figure 1: System structure diagram.

Because the signals collected by the sensors are in different forms, the signals need to be processed effectively, and then used to the interrupt mode is passed to the CPU, and the corresponding processing is made, and then the control signal is transmitted.

The server adopts the WAMP integrated environment, and the database mainly contains two tables of user and music. Web-side and server-side data exchange is through HTTP, and the communication is directly carried out through the website URL. The navigation between pages is passed by adding "?+variable name=variable value" after the url address of the webpage, and then the system array GET is used for extraction.

3 KEY TECHNOLOGY

The technical solution adopted in this article is: a voice interaction system based on gesture recognition, as shown in Figure 2 and Figure 3, which includes a distance sensor, a central processing unit, a camera, a microphone, a storage module, a vibration motor, and a communication module. The distance sensor, the camera, the microphone, the storage module, the vibration motor and the communication module are respectively electrically connected to the central processing unit. The distance sensor is used to detect the distance between the camera and the hand of the target object,

and the camera is used to The hand senses and collects the gesture action image of the target object, the camera sends the gesture action image of the target object to the central processing unit; the microphone is used to collect voice instructions of the target object; the storage module is provided with gesture storage Area, voice pre-command storage area, temporary voice storage area and mixed voice storage area. The gesture storage area is used to access preset gesture actions. The temporary voice storage area is used to store first voice instructions. A voice command is a voice command of a target object collected by a microphone, the voice pre-command storage area is used to store a second voice command, the second voice command is a voice command corresponding to a preset gesture action, and the mixed voice storage area is used for Store a third voice instruction, the third voice instruction is formed by splicing and combining the first voice instruction and the second voice instruction; the central processing unit is used to perform gesture recognition processing on the gesture action image of the target object collected by the camera to obtain the camera The preset gesture action corresponding to the collected gesture action image of the target object.

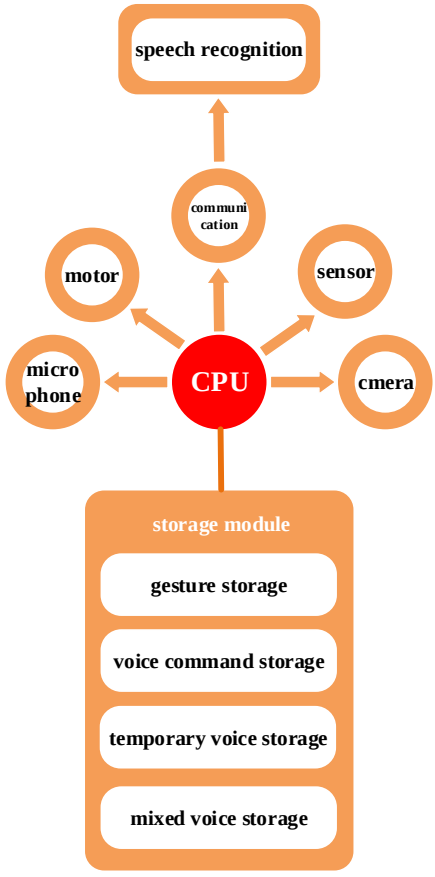

Figure 2: Voice interaction system.

The voice interaction system based on gesture recognition is connected to an external electronic device through a communication module. The central processor sends the voice command in the storage module to the external electronic device through the communication module. The communication module is Bluetooth, wifi or other wireless connection modules.

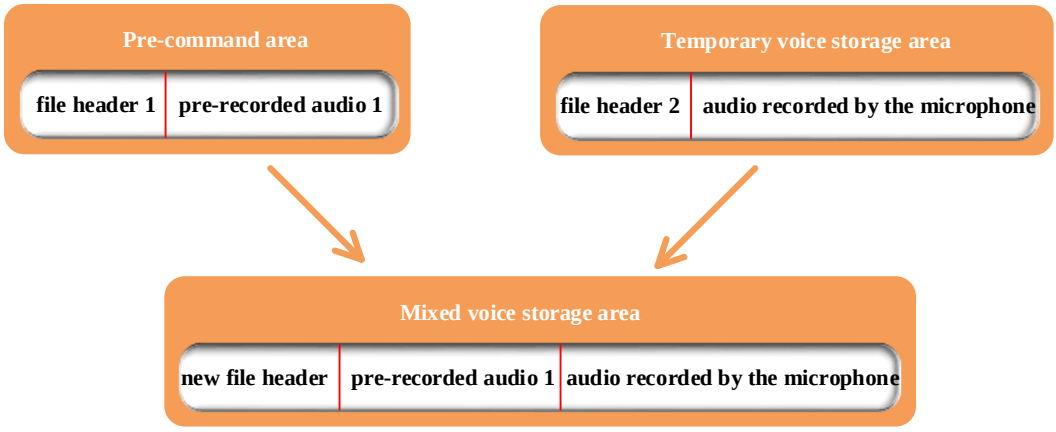

Figure 3: Storage method.

Make the system connect with external electronic devices through Bluetooth, wifi or other wireless connection methods. The distance sensor is used to detect the distance between the camera and the target object's hand, and compare the detection distance with the preset distance. When the distance between the camera and the target object's hand is within the preset distance range, move toward the center The processor sends a trigger signal; the central processor controls the camera to start working in response to the trigger signal sent by the distance sensor, wherein the target object is preferably a user. The predicted results are shown in Figure 4 and Figure 5.

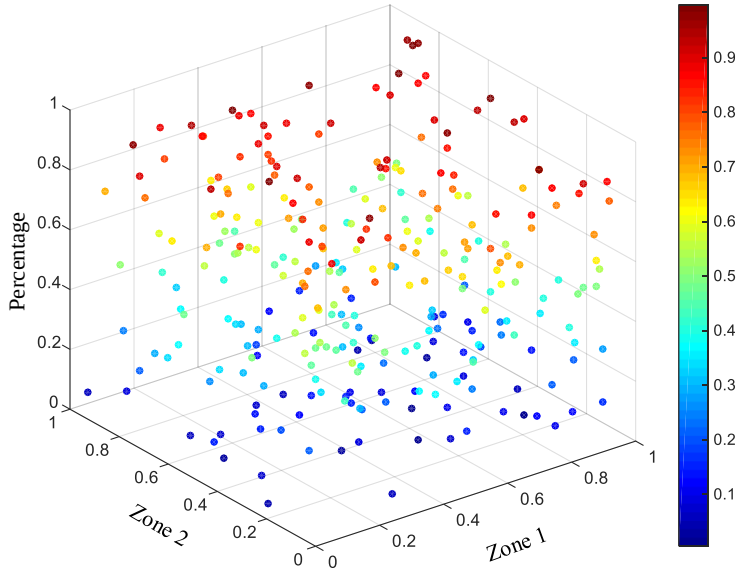

Figure 4: Percentage at various zone.

The distance sensor detects that the distance between the camera and the target object's hand is within a preset distance range for a time period exceeding a preset threshold, and sends a trigger signal to the central processing unit. The camera is used to sense the hand of the target object and collect the gesture action image of the target object.

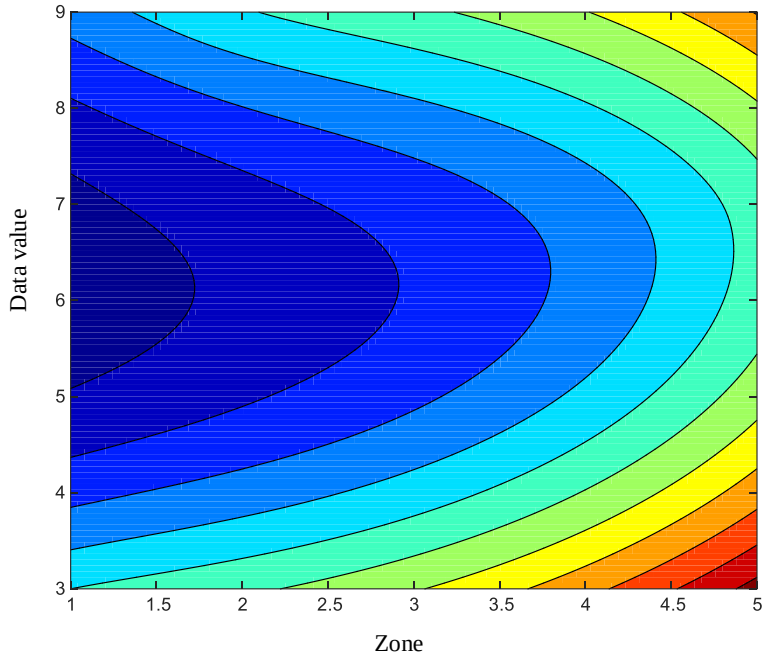

Figure 5: Value of data.

The camera sends the gesture action image of the target object to the central processing unit; the microphone is used to collect voice instructions of the target object. The storage module is provided with a gesture storage area, a voice pre-command storage area, a temporary voice storage area, and a mixed voice storage area. The gesture storage area is used to access preset gesture actions, and the temporary voice storage area is used to store the first A voice command, the first voice command is a voice command of a target object collected by a microphone, the voice pre-command storage area is used to store a second voice command, the second voice command is a voice command corresponding to a preset gesture action, The mixed voice storage area is used to store a third voice instruction, the third voice instruction is formed by splicing and combining the first voice instruction and the second voice instruction, and the central processing unit stores the third voice in the mixed voice storage area. The instruction is sent to the external electronic device through the communication module. Specifically, the voice pre-instruction storage area stores a plurality of second voice instructions in wav format, wherein the voice pre-instruction storage area is provided with a plurality of small partitions, such as a pre-instruction area, a pre-instruction area, and a pre-instruction area. There are three command areas, each of which contains a second voice command in wav format, and each second voice command in wav format is in a one-to-one correspondence with each preset gesture action in the gesture storage area; the temporary voice storage The first voice instruction in the wav format is stored in the area, the second voice instruction in the wav format stored in the voice pre-instruction storage area and the sampling number and sampling number of the first voice instruction in the wav format stored in the temporary voice storage area The settings of frequency and number of channels are the same. The results are compared in Figure 6 and Figure 7.

The central processing unit is configured to perform gesture recognition processing on the gesture action image of the target object collected by the camera, and obtain the preset gesture action corresponding to the gesture action image of the target object collected by the camera; specifically, the central processing unit takes the target object collected by the camera.

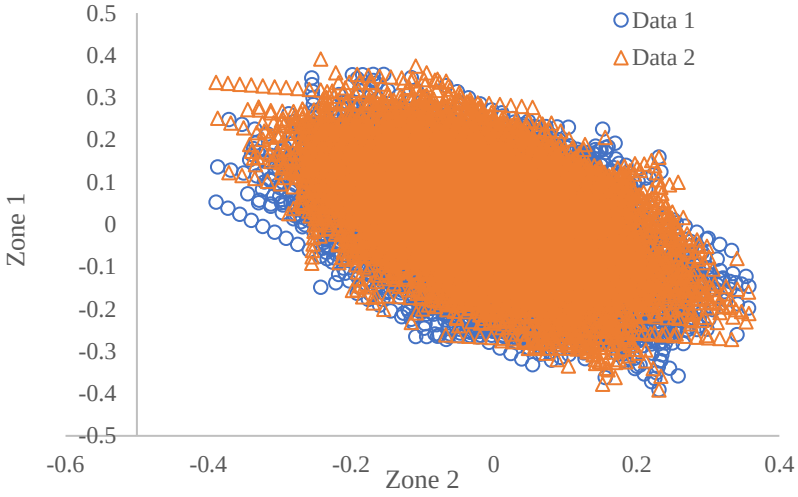

Figure 6: Data in different zone.

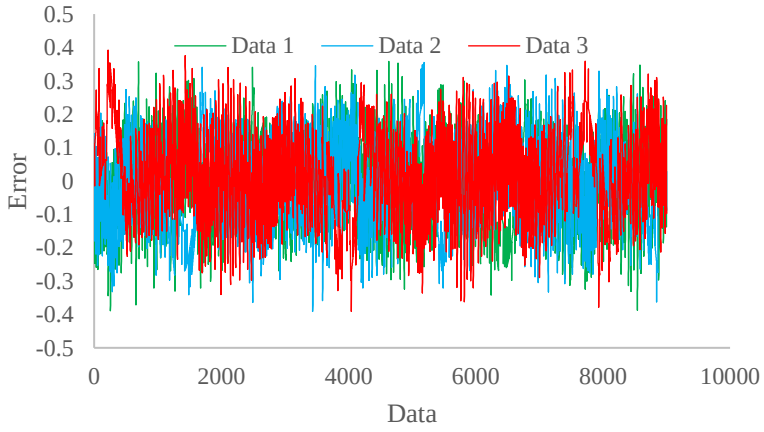

Figure 7: Error.

The gesture image in the gesture action image is separated, the features are extracted and compared with the preset gesture action, so as to determine the preset gesture action corresponding to the gesture action image of the target object collected by the camera; when the collected gesture action image of the target object is When the gesture action is preset, the central processing unit sends a control signal to the vibration motor and the microphone, the microphone starts to collect voice instructions from the target object, and the vibration motor starts to operate to indicate that the gesture recognition processing is successful; wherein the gesture recognition processing method is The techniques commonly used by those skilled in the art will not be repeated here. When the gesture motion image of the target object collected by the camera and the preset gesture motion continue to match successfully, it is determined that the target object is "holding hand gesture", the microphone continues to collect the voice command of the target object, and the vibration motor continues to vibrate; when the camera When the collected gesture action image of the target object cannot be successfully matched with the preset gesture action, it is determined that the target object is "no longer hold the gesture of raising the hand", the microphone stops collecting the voice command of the target object, and the vibration motor stops vibrating. When the gesture recognition processing is successful, the central processor determines that the gesture image of the target object

collected by the camera matches a certain preset gesture action information in the gesture storage area, and the central processor sends the corresponding pre-instruction voice in the voice pre-command storage area. The second voice command is read, and the audio file corresponding to the second voice command is spliced with the audio file corresponding to the first voice command in the temporary voice storage area to form an audio file corresponding to the third voice command. Specifically, an audio file in the wav format contains a file header part and an audio information part. The bytes of the file header record the identifier of the audio file in the wav format, the data length of the audio information, and the number of sampling bits of the audio information. Setting of sampling frequency and number of channels; when splicing the audio file corresponding to the second voice command with the audio file corresponding to the first voice command, first delete the file headers of the wav audio files of both, and then according to the two Setting information of the total length of audio data, number of sampling bits, sampling frequency, and number of channels, a new file header is regenerated, and then the new file header, the first The audio file corresponding to the second voice command removes the original file header, and the audio file corresponding to the first voice command removes the original file header and recombines to form a new wav audio file, that is, the audio file corresponding to the third voice command, which is The common methods used by those skilled in the art need not be repeated here.

One song, changing the background music mode, etc., so the voice recognition technology mainly adopts the template matching method. The method has four main steps: feature extraction, template training, template classification, and judgment. Because if a codebook is optimized for a specific information source, the average quantization distortion of the signal generated by this information source and the codebook should be less than the average quantization distortion of other information signals and the codebook , Which means that the encoder itself has the distinguishing ability.

4 CONCLUSION

The development of this system makes full use of the API provided by the open source operating system Linux and the efficient programming language C to develop. It not only adapts to the limited software and hardware resources of the embedded system, but also increases the running speed of the program. However, the implementation of this system still needs to be improved. Due to the limited majors, the hardware and drivers in the experiment are purchased or customized. What we achieve is mainly the algorithm of the central control system, the design of the communication protocol with the sensor, and the website Design, etc. At present, the system is able to run successfully and has been simply deployed. Future work will be further in-depth, mainly to be able to use mobile phones to intelligently remotely sense music playback, improve the system, improve the design of the database, increase the collection of sensors, and perform more accurate playback of the most suitable songs for users.

5 ACKNOWLEDGEMENT

Henan Province Educational Science "14th Five-Year Plan" General Project for 2021 "Research on the Status Quo and Strategies of Music Education and Teaching in Higher Normal Education in China" (Project No. 2021YB0626)

Jun Zhao, https://orcid.org/0000-0002-0475-8383
Marianne Zhao, https://orcid.org/0000-0001-8809-6841

REFERENCES

[1] Kang, D.; Seo, S.: Personalized smart home audio system with automatic music selection based on emotion, Multimedia Tools and Applications, 78(3), 2019, 2367-2376. https://doi.org/10.1007/s11042-018-6733-7

[2] Gómez, J.; Moro, L.; Godino, J.: On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors, Biomedical Signal Processing and Control, 8(31), 2019, 101-107. https://doi.org/10.1016/j.bspc.2018.09.003
[3] Baris, B.; Ioannis, G.; Yannis, S.: A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection, Computers in Biology and Medicine, 100, 2018, 132-143. https://doi.org/10.1016/j.compbiomed.2018.06.026
[4] Keivanpour, S.; Kadi, D.: Internet of Things Enabled Real-Time Sustainable End-of-Life Product Recovery, IFAC PapersOnLine, 13(52), 2019, 796-801.https://doi.org/10.1016/j.ifacol.2019.11.213
[5] Lee, J.-E.; Hur, S.; Watkins, B.: Visual communication of luxury fashion brands on social media: effects of visual complexity and brand familiarity, Journal of Brand Management, 25(5), 2018, 449-462. https://doi.org/10.1057/s41262-018-0092-6
[6] Wenjuan, L.: The Integration of Contemporary Art Visual Elements in Visual Communication Design, Journal of Frontiers in Art Research, 1(3), 2021, 4-7. https://doi.org/10.23977/jfar.2021.010302
[7] Fan, M.; Li, Y.: The application of computer graphics processing in visual communication design, Journal of Intelligent & Fuzzy Systems, 39(4), 2020, 5183-5191. https://doi.org/10.3233/JIFS-189003
[8] Kaleli, Y.-S.: The Effect of Computer-Assisted Instruction on Piano Education: An Experimental Study with Pre-Service Music Teachers, International Journal of Technology in Education and Science, 4(3), 2020, 235-246. https://doi.org/ 10.46328/ijtes.v4i3.115
[9] Gilbert, T.: Looking at Digital Art: Towards a Visual Methodology for Digital Sociology, The American Sociologist, 49(4), 2018, 569-579. https://doi.org/10.1007/s12108-018-9384-2
[10] Pérez, G.; Manuel, T.-J.; Morant, R.: Cantus: Construction and evaluation of a software solution for real-time vocal music training and musical intonation assessment, Journal of Music Technology and Education, 9(2), 2016, 125-144. https://doi.org/ 10.1386/jmte.9.2.125_1