Audio Recognition and Interactive Generation System through Digital Information Identification and Artificial Intelligence Technology

Yuxiao Zhang1, 2, *
[1.] International School, Beijing University of Posts and Telecommunications, Beijing, China, zhangyuxiao@bupt.edu.cn, *Corresponding author; [2.] Artificial Intelligence School, Beijing University of Posts and Telecommunications, Beijing, China

Yujie Yin1, 2
[1.] International School, Beijing University of Posts and Telecommunications, Beijing, China; [2.] Artificial Intelligence School, Beijing University of Posts and Telecommunications, Beijing, China

Yue Wang
Department of Composition, China Conservatory of Music, Beijing, China

ABSTRACT

Music information extraction has been a popular research direction in artificial intelligence technology, which can identify the digital information of the music in the audio, including the passage, timbre, and track of the music, based on the audio files uploaded by users. This paper applies the recognition of deep learning framework to music recognition and music information extraction. Based on this deep learning framework, a music recognition and interactive generation system based on artificial intelligence technology is designed and implemented using server-side and client-side development techniques. This system can recognize and generate audio project files uploaded by users, and at the same time, users can use the interactive music generation system to edit and create this audio project file, and finally generate audio files based on the modified audio project. This overall system, on the one hand, lowers the threshold for users to create music, and on the other hand, provides great convenience for music creators.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**;

KEYWORDS

Music information extraction, Music recognition, Deep learning framework, Audio engineering, Interactive

ACM Reference Format:
Yuxiao Zhang1, 2, *, Yujie Yin1, 2, and Yue Wang. 2021. Audio Recognition and Interactive Generation System through Digital Information Identification and Artificial Intelligence Technology. In *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM2021), October 23–25, 2021, Manchester, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3495018.3501090

1 INTRODUCTION

With the development of artificial intelligence and the popularity of electronic devices, the digital processing and recognition of audio signals such as voice and music has provided great convenience for arranging music, showing a broad market prospect. Based on the feedback from users, interactive arranging systems have been gradually iterated and their functions have been expanded while helping the development of contemporary music. This paper combines artificial intelligence technology with electronic interactive arranging system to propose a new type of interactive system for arranging music, which can identify the style, style, and timbre track of the audio for the user based on the user's simple audio, such as humming, and then hand over the whole audio project to the user for processing, and the user makes electronic interactive adjustments to the details of the audio process to finally complete the whole The work is finally completed. The running process of the system includes the user uploading audio files or using the recording module on the client side, the server side performing audio recognition, generating the initial audio project file, transferring the audio project file to the client side, the user modifying the audio project on the client side, and finally performing music generation. The operation flow of the system is shown in Figure 1

2 MUSIC RECOGNITION SUBSYSTEM BASED ON ARTIFICIAL INTELLIGENCE TECHNOLOGY

2.1 General framework of subsystem

The artificial intelligence-based piano arrangement timbre recognition system designed in this paper is mainly composed of training sample set, timbre mining, timbre recognition and other parts. The overall framework diagram of the system is shown in Figure 2
The general framework of the music recognition subsystem with artificial intelligence technology consists of the following parts.
(1) Audio database. It includes recognition training dataset and test dataset. The original audio of the song constitutes the training data set. The test data set contains not only a part of the above mentioned audio, but also a part of the humming audio data of non-professionals, in order to compare and evaluate the generalization ability of the model.

AIAM2021, October 23–25, 2021, Manchester, United Kingdom
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8504-6/21/10…$15.00
https://doi.org/10.1145/3495018.3501090

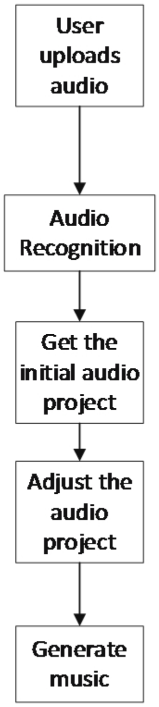

Figure 1: Music Generation System.

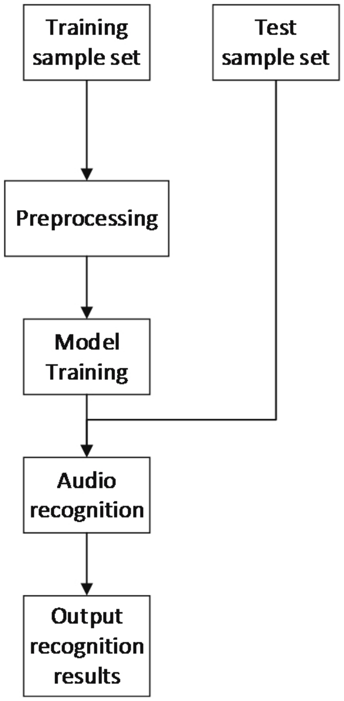

Figure 2: General framework of music recognition subsystem.

(2) Pre-processing module. The humming audio data are audio processed and analyzed to obtain the audio data feature representation.
(3) Model training module. The training data set is input into the audio recognition neural network, and the training is carried out in batches. The loss function value is calculated after each iteration using the validation set, and the training is stopped after reaching certain accuracy requirements, and the weight value of the neural network with the smallest loss function value is adopted as the optimal parameter output of the model.
(4) Neural network testing module. Based on a certain amount of test audio data, suitable evaluation indexes are used to test and evaluate the neural network performance as the basis for repeatedly adjusting the neural network training process and parameter selection.
The use of modular design makes the audio recognition deep learning framework well scalable and yields several highly cohesive and low-coupling modules. For example, the audio database can be flexibly replaced or modified for neural network model training so that training can be evaluated on different datasets; a set of model parameters trained by the neural network training module can also be used to test network performance and end-to-end testing of the prototype system on different datasets.

2.2 Recurrent neural networks

The hum recognition neural network model is the core part of the deep learning framework, and its general design idea is as follows.
(1) The input layer receives as input the Mel spectrogram of the humming audio signal.
(2) Then several convolutional layers are used to learn the local features of the audio signal to get the audio signal feature mapping.
(3) Then several recurrent layers are used to generalize and learn the sequence features composed of the audio signal over time.
(4) Finally, the probability distribution of the recognition result for a song with the input audio signal is derived by the Softmax activation function.
According to the above design idea, the audio recognition neural network model is shown in Figure 3
Among it, the data received in the input layer is a two-dimensional Meer spectrogram representation of the audio signal. In order to fully extract the spectral features of the audio signal, multiple had convolution and pooling layers are used to learn the local features of the signal, and a threshold control unit is used in the last two layers to learn the sequence features of the audio signal for generalization. Finally, a Softmax activation function is used in the output layer to output the neural network computation results as a probability distribution for song recognition.
To fully learn the local features of the audio signal, the model uses multiple layers of convolution and pooling layers. The combination of multiple convolutional layers is able to approximate complex nonlinear model parameters while still allowing training learning in a relatively simple way. In fact the range of model functions that can be represented by a network composed of multiple convolutional layers is the same as the range represented by a single convolutional layer, and the multilayer convolutional structure with nonlinear superposition can greatly extend the range of functions that can be represented by a neural network.

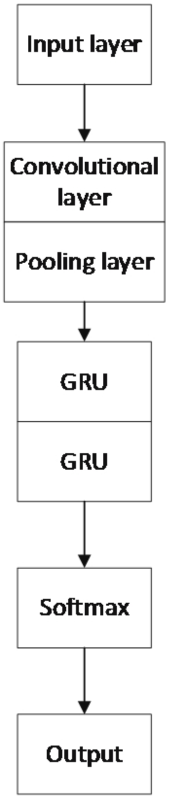

Figure 3: Neural network model for audio recognition.

From the definition of convolution operation, it is clear that the output result size will become smaller after convolution operation without filling the input data. A pyramid-like structure can thus be formed, where the feature mapping becomes smaller in size as the level deepens, but the number of channels becomes larger and larger.

At the shallower levels, the convolution extracts more detailed features, such as the pitch of the note, the time value of the note, etc. And at the deeper levels, the previous detailed features can be extracted and combined twice, which is equivalent to looking at a larger field of view, thus extracting more complete and abstract features, such as a melody. Ultimately, the resulting output can be much better than processing the raw audio directly.

2.3 Interaction design of the subsystem networks

The audio sheet music sub-recognition system adopts C/S architecture, and the client and the server communicate through HTTP protocol. The interaction design flow is shown in Figure 4

Users input audio signals or upload audio files through the client-side recording module, and the client-side sends HTTP POST requests through the audio upload module to upload audio to the server-side for processing. The server-side audio pre-processing module processes the audio for note onset detection, audio segmentation, etc., and then inputs it to the recognition module, which uses the audio recognition neural network model to give recognition results and returns the corresponding audio project file for the client to display.

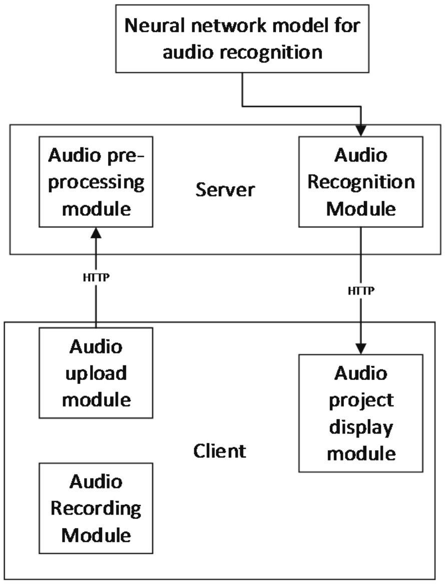
Neural network model for audio recognition
Audio pre-processing module
Server
Audio Recognition Module
HTTP
HTTP
Audio upload module
Client
Audio project display module
Audio Recording Module

Figure 4: Interaction Architecture Diagram.

3 INTERACTIVE MUSIC GENERATION SUBSYSTEM

3.1 Subsystem operation flow

In the music recognition subsystem, the audio project file of the user's uploaded audio is obtained. In this project file, the style of the music is firstly determined, which makes the rules for dividing the whole project into paragraphs, and the music is divided into different paragraphs, each paragraph contains different tones, and each tone corresponds to a track and segment. Users of the system can edit and audition the different tracks and clips in the project according to the style of the music, and get a perfected audio project file, which is finally packaged to produce an audio file such as MP3 format. The operation flow of the interactive music generation subsystem is shown in Figure 5

3.2 Subsystem functional module design

The specific functional modules of the interactive music generation subsystem can be divided into audio simulation module, audio editing module and score display module.

In the audio simulation module, the whole audio or fragment can be played, and the digitized audio engineering data can be converted into analog signals, so that people can feel the editing results of the project and compare the audio differences before and after editing from time to time.

In the audio editing module, on the one hand, you can artificially remove the interference audio or noise generated in the process of audio recognition, and on the other hand, you can recreate the results of audio recognition, modify the sound track timbre of each

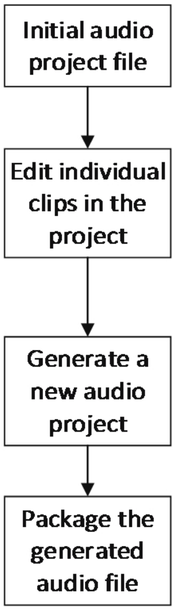
Initial audio project file
Edit individual clips in the project
Generate a new audio project
Package the generated audio file

Figure 5: Interactive music generation subsystem operation flow.

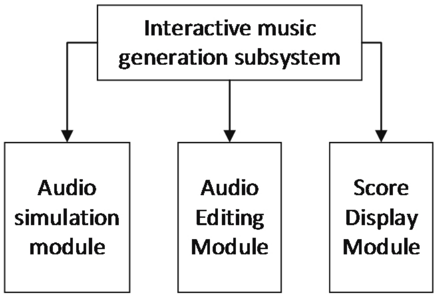

Figure 6: Sub-system functional module design.

paragraph in the music, making the elusive music inspiration into a reality and data that can be recorded.

In the score display module, the user can display and generate the score for the current edited audio project, so that the whole audio project can be reproduced manually or electronically and repeatedly played.

The subsystem functional module design is shown in Figure 6

In the interactive music generation subsystem, even people with no musical knowledge can hum a tune and then edit this for audio engineering to generate music of their own composition.

4 CONCLUSION

This paper proposes a deep learning framework for audio recognition based on artificial intelligence technology, introduces deep learning methods to the audio recognition problem, combines traditional signal processing methods and deep learning methods, provides new ideas for audio recognition technology research, and provides theoretical basis and processing methods for transforming audio signals into deep neural network input vectors. Meanwhile, this paper designs and implements a deep learning-based audio recognition system using server-side and client-side development techniques, and designs an interactive music generation system based on this audio recognition system, in which users can edit and create music with only simple audio files, which can greatly reduce the threshold of music creation and improve the creation efficiency of music creators This system can greatly reduce the threshold of music creation, and at the same time increase the efficiency of music creators, enabling the fusion of different styles of music and helping the development of music.

REFERENCES

[1] Zuo, Zhang, Chi. Research on piano timbre recognition and electronic synthesis system based on Fourier analysis method [J]. Automation Technology and Applications, 2021, 40 (02): 137-140+147.

[2] Tong Zhibei. Artificial intelligence-based piano arrangement timbre recognition system design [J]. Modern Electronic Technology, 2020, 43 (04): 183-186.

[3] Tian Peng. Research and prototype implementation of music style recognition and generation technology based on deep learning [D]. University of Electronic Science and Technology, 2019.

[4] Yu Yuanyuan. Music recognition based on deep networks and hash learning [D]. Nanjing University of Information Engineering, 2018.

[5] Fei Yuquan. Research on music fountain based on music feature recognition [D]. Changsha University of Technology, 2017.

[6] Cheng Meifang. Design and implementation of piano timbre recognition and electronic synthesis system [D]. University of Electronic Science and Technology, 2014.

[7] Chen Xu. Content-based audio hum recognition and retrieval system [D]. Shanghai Jiaotong University, 2008.