www.cad-journal.net
Computer-Aided Design™
AND APPLICATIONS

Synthesis and Optimization of Instrumental Tone Quality Using Deep Learning

Jinxing Mu[1] and Jie Tian[2]

[1,2] School of Education and Music, Hainan Vocational University of Science and Technology, Hainan Haikou 571126, China, [1]cmyyy@163.com, [2]music.tian_bj@zit.edu.cn

Corresponding author: Jie Tian, music.tian_bj@zit.edu.cn

Abstract. Traditional musical instrument performance and sound production methods have made it difficult to meet the diverse needs of modern music creation. This article aims to improve the synthesis and optimization technology of tone quality by using a deep-learning algorithm. This article effectively constructs a melody model for the complex music structure of an independent encoder using the baseline of a music converter. By analyzing models with different baselines, the chord encoding part of the music structure was constructed. We constructed sound recognition by transforming and generating music under complex conditions. The results indicate that the proposed model has higher and richer key performance in the coverage of sound chords. Its model benchmark in graphic commerce has been expressed in a more diverse way. The histogram and pitch have been better reflected in the richness of performance processing. In terms of the instrumental value of music diversity, it demonstrates a highly optimized harmony.

Keywords: Deep Learning Algorithm; CAD Technology; Tone quality Synthesis; Music Generation Model; Sound Optimization
DOI: https://doi.org/10.14733/cadaps.2025.S1.283-296

1 INTRODUCTION

With the continuous development of society and the improvement of living standards, people's pursuit of music is also increasing. Musical instruments have become a great tool for people to relax and have fun after work. And electronic instruments, with their unique working mode, can perfectly combine traditional instruments with modern digital processing technology. Electronic instruments are increasingly being accepted and valued by the public due to their ability to provide more entertainment functions. In today's life, more and more entertainment devices are becoming familiar and accepted by the public. As the core tool of music creation, the musical instruments' sound characteristics and expressive force directly affect the emotional expression and artistic style of music works. In response to the serious loss of Chinese folk dance culture, the high cost of traditional instrument sound synthesis methods, and the high demand for professional background, Cai et al. [1] proposed a comprehensive method for the automatic generation of folk dance movements and instrument sound synthesis. This model can simulate the timbre and performance techniques of traditional instruments, and generate high-quality instrument performance audio for a given music segment. In terms of dance action generation, it has constructed a sequence-to-sequence network model. Through this approach, not only can Chinese folk dance culture be effectively inherited and promoted, but also the cost of manual choreography can be reduced, allowing more people to enjoy the charm of folk dance [2]. By using advanced feature extraction tools and multi-scale fused high-resolution networks, we extract their respective features from music, dance, and instrument audio. In terms of instrument sound synthesis, it utilizes features extracted from instrument audio and combines deep learning techniques to develop a new instrument sound synthesis model. However, the traditional sound production method is limited by physical conditions and human factors, and it is difficult to accurately simulate and reshape the sound characteristics of various musical instruments. In this experience, sound is not only a core element in building an immersive virtual world but also a key starting point for creative and technological exploration.

Throughout all electronic instrument products currently designed in China, there is a requirement to improve timbre and sound quality. Whether it is an electronic keyboard, electric piano, or electronic drum, further improvement of sound quality is needed. However, there are still significant limitations in improving sound quality and enhancing the quality and quantity of effectors currently used in China (due to major manufacturers such as Roland,...) The core chips used by YAMAHA are all their own and cannot be purchased on the market. As for most electronic instrument systems currently designed in China, there are only system effectors, and the effects are very poor. It is difficult to add some effects (i.e., insert effects) to the sound, and even if individual insert effects can be added, the quality is relatively poor. Similarly, it is even more impossible to separate the effects of music, keyboard, or drum (some advanced electronic instruments can do this). In addition, there are significant limitations on the number of pronunciations, as 64 simultaneous pronunciations are no longer sufficient to meet the current requirements for simultaneous pronunciations in electronic instruments. The current chip platform has significant limitations on further improving sound quality and music quality. This requires research on more high-end sound source chips and platforms to meet the requirements of system product design. Dong et al. [3] analyzed and compared commonly used audio retrieval methods and designed and implemented an abnormal sound monitoring method based on the spectral characteristics of sound. Due to the fact that abnormal sound data is often much smaller than scene sound data during data collection, the proportion of various types of samples is imbalanced, and it is necessary to balance various types of samples. Train audio feature parameters using a random forest model to obtain the final predicted category of the input audio signal and achieve monitoring of abnormal sounds. In order to verify the performance and cost of the algorithm, experiments were conducted on the collected dataset containing abnormal sounds and scene sounds. By comparing and analyzing the experimental results, it has been proven that this method has good accuracy and real-time performance and can still maintain good performance in strong noise backgrounds. Based on the above methods, an abnormal sound monitoring system based on sound recognition technology has been developed. This mainly includes a sound monitoring function, a real-time query function of sensor status, a visualization function of pickup data, and a storage and query function for historical data pickup. A systematic application analysis was conducted under different signal-to-noise ratios, using tunnel water inrush accidents as the application background. The results show that the system has a strong ability to recognize water inrush sounds and resist noise, meeting the requirements of timeliness, and can provide a basis for monitoring water inrush accidents. The automatic classification of music is not only a simple speech signal recognition problem but also closely related to instrument sound synthesis, both of which occupy important positions in the field of music information processing. The synthesis of instrument sound usually involves the analysis and simulation of the spectral, timbre, dynamic, and other characteristics of instrument sound. These characteristics are closely related to feature analysis in music classification. Ge et al. [4] conducted in-depth research on music classification methods based on feature analysis and also combined the principles and techniques of instrument sound synthesis to systematically explore the principles, methods, and techniques of music classification. Therefore, it attempts to apply sound analysis techniques in instrument sound synthesis to feature extraction in music classification, in order to improve the accuracy of music classification. It integrates timbre analysis and spectrum analysis techniques from instrument sound synthesis into the feature extraction process of music classification. This method can fully utilize the advantages of different feature analysis methods to improve the accuracy and robustness of music classification. Through in-depth analysis of timbre and spectral features, key information in music can be more accurately captured, such as instrument type, music style, emotional expression, etc. Through CAD technology, designers can accurately construct the three-dimensional model of musical instruments, and carry out acoustic simulation and optimization design. This digital-based design method greatly improves design efficiency and quality and enables designers to better understand and master the acoustic characteristics of musical instruments. Georges and Seckin [5] explored information visualization techniques applied to classical composer databases, and further combined tone quality synthesis techniques to reveal the correlation between classical music and contemporary music creation. In an era filled with prescribed lists of composers and works, or automated recommendation algorithms, our analysis provides an alternative path that may promote active discovery of composers and their music in a more free and creative way, and stimulate exploration and experimentation in instrument sound synthesis. It further combines data on the influence of style, the "ecology" of composers, and data on instrument sound synthesis to construct a comprehensive framework for analyzing the impact of composers on their music. In order to present these connections more intuitively, it used multidimensional scaling analysis techniques to locate composers on a map while preserving their paired distances. In this way, we can clearly see the similarities and differences among composers, as well as how these similarities and differences affect the selection and synthesis of instrument sounds. In the aspect of tone quality synthesis and optimization, CAD technology can provide rich data support for deep learning models, including the geometric structure, material properties and vibration characteristics of musical instruments. Wind instruments produce sound through the vibration of airflow, and their timbre is influenced by the shape, length, and playing method of the pipe. When delving into the Orchid Orchestral Quality (OOQ) framework, we inevitably need to address the importance of instrument sound synthesis and its potential connection with the OOQ framework. The OOQ framework provides composers with a new perspective by adopting analogies with digital signal processing, allowing them to more flexibly manipulate music materials, thereby enhancing or reducing their specific quality. Ghisi and Cella [6] adjust the parameters in the OOQ framework, such as frequency, amplitude, waveform, etc., to change the timbre, volume, and sound quality characteristics of music materials. Different instruments, due to their unique physical structure and sound production mechanism, produce their own distinctive timbre and sound quality. The synthesis of instrument sound is a crucial step in music production, which involves a deep understanding and simulation of the characteristics of instrument sound. These data are very important for the training and optimization of deep learning models, which is helpful to further improve the accuracy and quality of sound synthesis. Han [7] used ERP technology to design a human-machine interaction interface layout optimization framework for instrument sound synthesis software. This method combines user habits and cognitive characteristics to ensure that the optimization of interface layout matches the actual needs of users. They use the G_1 method that supports architecture to determine the importance of human-computer interaction interfaces in facing layout goals. By reasonably laying out interface elements, it is hoped to guide users' visual attention and enable them to complete sound synthesis operations more quickly and accurately. The experimental results show that by applying our proposed ERP-based instrument sound synthesis software human-computer interaction interface layout optimization method, user satisfaction has been significantly improved. ERP technology can capture the brain activity of users while operating the interface, providing a scientific basis for optimizing interface layout. Less than 0.3% of users expressed dissatisfaction with the system, which proves the advantage of this method in terms of user satisfaction.

In the research of music emotion recognition (MER) and instrument sound synthesis, researchers often rely on supervised learning methods based on music features and corresponding annotations. He and Ferguson [8] proposed a two-stage model based on segmentation that combines unsupervised learning and supervised learning, emphasizing the potential value of instrument sound synthesis in music emotion recognition. These fragments not only contain melody and rhythm information of the music but may also include specific instrument sounds and performance techniques. This step helps us capture the characteristics of different instrument sounds in music and their interactions, providing richer information for subsequent music emotion recognition. In the second stage, we use these unsupervised learned music segment features as inputs for the bidirectional long short-term memory deep learning model. Due to the use of fragment-based inputs, the model is able to focus on local details in music, while also increasing the size of training samples, which helps to reduce the risk of overfitting in deep learning processes. By randomly masking a portion of frequency components or time segments, we force the model to learn and recover complete music segments from the remaining information, thereby enhancing the model's robustness and generalization ability. Although deep learning algorithms and CAD technology have great potential and advantages in tone quality synthesis and optimization, the current research and application still face some challenges.

(1) In this study, the deep learning algorithm is combined with CAD technology for the synthesis and optimization of tone quality, which brings innovative technical integration to the field of music production.

(2) Through the training of deep learning model, the accurate simulation of different musical instruments is realized, which breaks the limitations of traditional sound production methods.

(3) The 3D modelling and acoustic simulation of musical instruments using CAD technology provide abundant data support for the deep learning model and enhance the accuracy and reliability of sound synthesis.

(4) The sound synthesis method based on deep learning has high flexibility and adjustability and can generate synthesized sounds with different styles and timbres as required.

The structure of this article is as follows: Firstly, in the introduction part, the importance of tone quality synthesis and optimization in music production and the potential application of deep learning algorithms and CAD technology in this field are expounded. Then, the basic principles of deep learning algorithm and CAD technology and their applications in sound synthesis and optimization are introduced through an overview of relevant theories and technologies. Then, the construction of a deep learning model of tone quality synthesis and optimization, and the method of musical instrument acoustic simulation and optimization based on CAD technology are discussed in detail. Then, the combination of the deep learning model and CAD technology is discussed, and the superiority of the proposed method is verified by experiments and results analysis. Finally, the research results are summarized in the conclusion and prospect part, and the limitations and improvement direction of the research are pointed out.

2 RELATED WORK

Huang et al. [9] proposed an end-to-end note detection model based on deep convolutional neural networks and feature fusion. Throughout the entire music production process, music object detection is a crucial part of the OMR pipeline. It is not limited to the recognition of sheet music, but also plays an important role in music production techniques such as instrument sound synthesis. Therefore, note recognition is not only the core and key of score recognition but also a crucial link in instrument sound synthesis technology. This model can directly process score images and effectively extract pitch, duration, and other related features of notes through deep learning and feature fusion techniques. This model has achieved significant performance improvement in note recognition, and we show its high-precision performance in general music symbol recognition tasks. By inputting parameters such as pitch, duration, and timbre of notes into a sound synthesis engine, realistic and expressive instrument sounds can be generated. The duration accuracy reached 0.92 and the pitch accuracy reached 0.96. These precise data not only demonstrate the effectiveness of the model in note recognition but also provide reliable data support for subsequent instrument sound synthesis. The application of digital media technology in various fields is increasingly deepening, not only promoting the development of various social classes but also opening new doors in the fields of art and innovation. Jiang [10] studied the digital media application technology of mobile terminals based on edge computing and virtual reality and paid special attention to the application of these technologies in musical instrument sound synthesis. The research used SD-CEN (software-defined central edge network) architecture and FWA (a specific optimization algorithm, such as a fuzzy weighting algorithm) to carry out simulation experiments in the edge computing environment. Compared with traditional cloud computing architectures, SD-CEN-based network architectures exhibit significant advantages in instrument sound synthesis. Especially in the field of instrument sound synthesis, the application of digital media technology provides unprecedented creative possibilities for musicians and sound designers. By using the methods of edge computing and virtual reality, audio signals can be captured and processed in real-time, and then the realistic sound of musical instruments can be synthesized. This architecture can optimize the allocation of network resources, ensure efficient processing of audio data at edge nodes, thereby reducing latency and improving the real-time and quality of synthesized sound. The results show that FWA performs outstandingly in reducing the response delay of real-time instrument sound synthesis services, and can significantly improve the user experience.

Klein et al. [11] proposed a cross-modal steady-state effect that not only reveals how music affects the inverted U-shaped relationship between visual perception complexity and people's liking but also delves into the unique role of instrument sound synthesis. When elements carefully crafted through instrument sound synthesis techniques are incorporated into the music, this influence is enhanced. This influence is not only reflected in changes in personal preferences but may also affect the audience's overall evaluation of visual advertising or products in the business environment. Specifically, we found that music (regardless of its complexity) typically shifts the optimal level of visual complexity towards people preferring relatively simple visual effects. When soft piano synthesized sounds are added to the background music, the audience may be more inclined to prefer visual designs with a minimalist and fresh style. These synthesized sounds can guide the audience's attention, thereby influencing their perception and preference for visual complexity. They remind us that when designing and presenting visual content, it is necessary to fully consider the influence of music, especially when specific instrument sound synthesis is incorporated into the music. Lossy audio codecs reduce file size by removing redundant information that is imperceptible to human hearing when compressing (and decompressing) digital audio streams. In addition to the audio enhancement and compression artefact removal, Lattner and Nistal [12] further explored the combination of instrument sound synthesis and audio restoration technology. During the training process, it pays special attention to the synthesis of instrument sounds. By introducing instrument sound libraries and specific synthesis algorithms, the generator can learn the unique features and timbre of instrument sounds. This random generator is conditional on highly compressed music audio signals, with the goal of producing outputs that are indistinguishable from instrument sounds in high-quality distribution versions. Research has found that compared to the 16 and 32kbit/s MP3 versions that only use traditional audio restoration techniques, a random generator combined with instrument sound synthesis can significantly improve the quality of audio signals, especially in terms of instrument sound clarity and timbre. However, high compression rates may introduce audio damage, which is particularly significant in the music field as they may affect the clarity and timbre of instrument sound.

Recommendation systems have been proven to be effective tools in predicting users' current music preferences. Liang and Willemsen [13] explored the development of user preferences in the field of instrument sound synthesis. Compared to users with lower music expertise, users with a deep music background exhibit higher preference consistency in selecting instrument sounds and synthesis effects. In instrument sound synthesis, users usually tend to choose tones and effects that are similar to their current preferences. However, further research is needed on how user preferences in instrument sound synthesis evolve over time and how these preferences affect the synthesis process. However, suppose the recommendation system can guide users to explore tones or effects that are slightly different from their current preferences, especially those that can stimulate new creativity. In that case, we may see users more open to experimentation and more innovative synthesis works. More representative default sliders, such as those recommending more representative genres of sound, may allow users to set sliders at a less personalized level, thereby limiting their creative space. Combining instrument sound synthesis with music visualization can create a brand-new music experience, helping people better understand and experience the charm of music and instrument sound. Liao [14] introduced the concept of music visualization and explored how to combine instrument sound synthesis with music visualization technology, as well as how to use deep learning algorithms to achieve this goal. Through this method, they associate the audio data synthesized from instrument sound with image data, thereby visually presenting the characteristics and changes of instrument sound. In order to fully utilize the emotional information in music and images, the visualization results can more accurately reflect the emotions expressed in music. By training models to simultaneously recognize emotional features in music and images, we can make music visualization results more vivid and infectious. The experimental results show that when the weight of the emotion classification loss function is set to 0.2, the improved deep learning-based music visualization algorithm has the highest matching accuracy.

Liao and Gui [15] proposed an innovative sparse feature extraction method based on sparse decomposition techniques and multiple instrument component dictionaries. This method first constructs a dictionary containing multiple instrument sound samples, which cover the sound characteristics of various instruments under different pitch, volume, and performance modes. By sparse decomposition of the input mixed instrument music signal, we can represent the signal as a sparse combination of elements in the dictionary, i.e., a sparse coefficient vector. Compared with existing methods, this method not only reduces the dependence on data labels but also goes beyond the limitations of only based on frequency domain or physical features, thus achieving significant performance improvement. This is because the synthesis of instrument sound is often closely related to the emotional expression of the performer, and sparse feature extraction methods can reveal the specific expression of this emotion in music signals. By analyzing these sparse coefficient vectors in depth, we can extract independent sparse music features. These features not only have high interpretability and can intuitively express the composition of musical instruments, but also accurately capture the changes in emotions in music [16]. By integrating the results of instrument sound synthesis into the visualization process, we can more clearly see the impact of different instrument sounds on the overall style of music. The clustering algorithm is used to discover similarities between different music works and classify them into different music genres or styles. In terms of audio editing, the system provides an intuitive and user-friendly interface, allowing users to easily crop, splice, and fade in and out audio. In terms of instrument sound synthesis, this system combines modern music production technology and computer technology to achieve simulation and synthesis of instrument sound [17]. This system can not only perform high-precision digital processing on recorded sounds but also present them intuitively on computer screens in the form of sound waves. At the same time, the status of all editing operations will be recorded in detail in the database, providing rich feedback and reference information for teaching and training. Students can create music in an intuitive and visual environment, deepening their understanding of music theory and methods through practical operations. It also integrates advanced instrument sound synthesis technology, providing a comprehensive and interactive teaching and learning platform for students and teachers [18].

3 SYNTHESIS AND OPTIMIZATION MODEL OF TONE QUALITY

3.1 Music Feature Analysis

In the realm of audio processing, a common assumption is that audio signals exhibit short-term stationarity, implying that within a brief temporal window (spanning approximately 20 to 40 milliseconds), the sinusoidal model's parameters remain largely stable. Leveraging this attribute, a short-duration audio signal can be viewed as a succession of nearly time-invariant sine waves. For audio signal gain adjustment, if the adjustment duration significantly exceeds the audio's natural variation cycle, say, 5 seconds, the gain alterations will tend to be gradual, minimizing abrupt fluctuations. This approach considers the audio's output characteristics to maintain processing stability.

However, in certain scenarios, like a marked rise in audio output intensity, prompt gain reduction is imperative to prevent equipment damage. This rapid adjustment mechanism is a vital component of audio processing. Moreover, when the audio output surpasses a predefined threshold, ensuring accuracy and responsiveness necessitates setting the gain adjustment time to a minimal 0.5 millisecond. Figure 1 graphically depicts sound source separation technology.

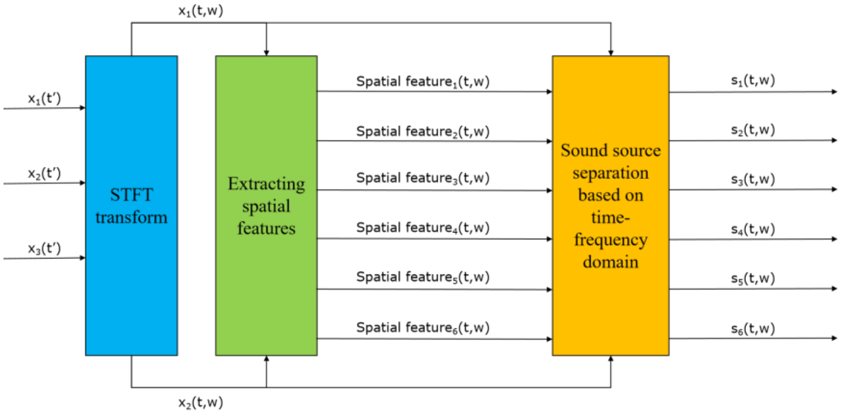

Figure 1: Schematic diagram of sound source separation technology.

The characteristics of sound mainly include four core dimensions: pitch, length, sound intensity and timbre, which correspond to the frequency, duration, amplitude and complexity of spectrum distribution of sound vibration respectively. Pitch is one of the most intuitive properties of sound, which is determined by the number of times the sound source vibrates in a specific time. When the number of times an object vibrates per unit of time increases, the generated pitch will increase accordingly, which is manifested by the increase of sound wave spacing; On the contrary, if the number of vibrations decreases, the pitch decreases and the sound wave spacing decreases. This article focuses on stringed instruments, especially taking the piano as an example, to deeply discuss these characteristic parameters of audio. As a typical stringed instrument, the piano's sounding mechanism is that the internal strings vibrate by tapping the keyboard, and then sound is produced. Each key of the piano corresponds to one or more strings. Striking different keys will trigger strings with different lengths to vibrate, thus producing sounds with different pitches. Furthermore, the duration of string vibration determines the length of the sound, and the amplitude affects the sound intensity, while the material of strings, production technology and the design of the resonance cavity jointly affect the performance of timbre.

Use the lateral propagation of sound waves to illustrate this process. The tension of the strings is T, the density is ρ, the cross-sectional area is A, and the length is 1. The strings are divided into $n+1$ segments with n the number of dividing points. The free vibration response of the strings is:

$$y\ \ x,t\ =\sum_{n=1}^{\infty}\left\{\frac{2}{1}\left[\int_0^1 f_1\ \ x\ \sin\frac{n\pi}{1}xdx\right]\cos p_n t+\frac{2}{nc\pi}\left[\int_0^1 f_2\ \ x\ \sin\frac{n\pi}{1}xdx\right]\sin p_n t\right\}\sin\frac{p_n}{c}x \tag{1}$$

The natural frequency of the system is:

$$p_n=\frac{n\pi}{1}\sqrt{\frac{T}{\rho A}}\quad n=1,2,3,\ldots \tag{2}$$

When $n=1$, p_n is called the fundamental frequency, which determines the pitch. It is found that there is an inverse relationship between the fundamental frequency of sound waves and density, and it is also inversely proportional to the length of the strings but directly proportional to the tension on the strings. Strings have different natural frequencies because of their different elasticity. When the rope is tightened, the pitch will rise accordingly, and vice versa. The total vibration form of strings is composed of the superposition of the amplitudes of many traveling waves. Among them, lateral vibration is the most basic vibration mode, and the decisive factor of timbre is segmented vibration mode, and amplitude directly affects the strength of sound. When an external force acts on the strings to deform them and make sounds, the generated sound waves will show different vibration amplitudes according to the different positions of the excitation points. Especially when the excitation point happens to be on the syllable, the homophonic will be obviously suppressed.

3.2 Musical Instrument Acoustic Simulation

The duration of a note signifies the extent of its sonic length, dictated by the time span of the vibrating pronunciation body. Conversely, pitch is primarily influenced by the rate of sound wave vibrations, with faster frequencies correlating to higher pitches and a more acute sound quality. Conversely, deep voices typically stem from lower pitch levels, a result of slower sound wave vibration frequencies.

In order to promote the learning of the model, the music should be preprocessed before it is encoded. Firstly, the MIDI input is quantized. Since most notes can be expressed by sixteenth notes, in the past, sixteenth notes were used as the time resolution in most music generation work in the symbolic domain. However, melodies that cannot be expressed by sixteenth notes, such as octave trio and thirty-second notes, have been abandoned in most studies at present, which will reduce the learnable characteristics of the model and reduce the learning effect of the model. Therefore, in this article, each input melody is quantized to 16 time steps in a bar, that is, 4 time steps per beat, and the sixteenth note is the minimum resolution of the duration. Figure 2 shows the process of quantizing irregular rhythm notes into sixteenth notes from two forms: symbol and piano roll.

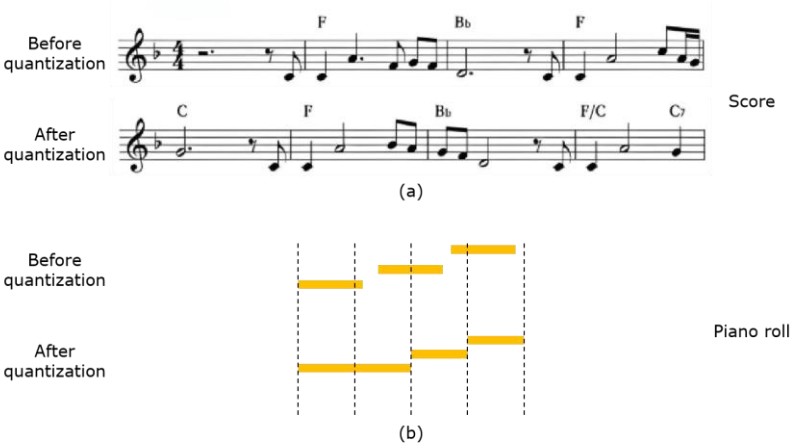

Figure 2: Quantitative comparison of score and piano roll.

In addition to quantization, the music should be tone-modulated. The distribution of key signatures in data sets is often uneven, and the pitch distribution of notes may be very different in different key signatures. In order to ensure the reliability of the experimental results, this article converts all the music into C major or A minor to make all the music consistent, so that the model can effectively learn the music patterns.

When the length is T , the melody sequence is:

$$m_{1,T}= \ m_1,m_2,\ldots,m_T \tag{3}$$

The chord sequence is:

$$c_{1,T}= \ c_1,c_2,\ldots,c_T \tag{4}$$

When the time step $t\in \ 1,2,\ldots,T$, the model generates four chord vectors c_t^{1st} , c_t^{2nd} , c_t^{3rd} and c_t^{1th} :

$$c_{1,T}= \ c_1,c_2,\ldots,c_T \tag{5}$$

Among them, M_c is a chord generation model based on $Transformer$ the model, θ_c its parameter, and $n\in \ 1,2,3,4$ the index of the chord vector. The chord sequence at c_t moment can be obtained by combining the four outputs with time step t .

In the realm of music signal processing, a diverse array of window functions are frequently employed, with the rectangular window and Hanning window serving as notable examples, alongside others:

Rectangular window:

$$w\ \ n\ \ =1,0\le n\le N-1 \tag{6}$$

Hanning window:

$$w\ \ n\ \ =0.5\left(1-\cos\left(2\pi\frac{n}{N-1}\right)\right),0\le n\le N-1 \tag{7}$$

When considering the frame length of a music signal, denoted as N it's crucial to understand that the choice of window function significantly impacts the analysis of its characteristic parameters.

$$\triangle f=\frac{1}{NT_s} \tag{8}$$

4 EXPERIMENT AND RESULT ANALYSIS

This model effectively solves the challenges faced by RNN in processing long-sequence music generation and has achieved remarkable results in reducing the complexity of spatial computation, which has a far-reaching impact on the field of music generation. On the other hand, Pop Music Transformer has been upgraded on the cornerstone of Music Transformer. By introducing a new representation, REMI, the generation effect of long-sequence music has been optimized. Because these two models have their own characteristics and use different representations, they are chosen as the baseline model of this article. In the model of this article, although the input sequence also covers the melody and chord of music, the processing method is different: the melody and chord are disassembled into independent sequences respectively, and then encoded by two encoders, and then input into the decoder. In order to verify the practicability of this study, we designed experiments and obtained the objective assessment index results of the baseline model and this model (Figure 3).

Figure 3 shows the comparison data of five assessment indexes between the chord and melody double coding model proposed in this article and the baseline model Music Transformer. The direction of the arrow in the table indicates the advantages and disadvantages of the numerical value: the higher the numerical value, the better, and the lower the numerical value, the better. The optimal results of each index are specially marked. The data show that the music generation model proposed in this article outperforms the baseline model in terms of chord coverage (CC), chord histogram entropy (CHE) and chord pitch distance (CTD). Specifically, the high values of CC and CHE reflect the richness of chord types, which means that this model can generate more diverse music. However, the improvement of chord diversity may sacrifice some smoothness, so the model in this article scores higher on the CTD index; that is, the smoothness of chord conversion is slightly lower than that of the baseline model.

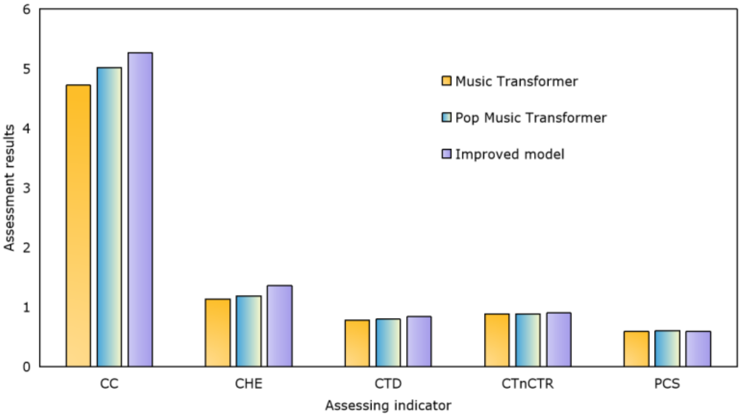

Figure 3: Objective assessment of music transformer and improved model.

The high score of the ratio of the chord to the inharmonic chord (CTnCTR) shows that there are fewer inharmonic chords in the music generated by this model, which is closer to the practice of artificial composition, thus enhancing the musicality. On the other hand, in terms of pitch and harmony score (PCS), because the baseline model adopts the training mode of combining melody and chord input, its harmony is better than the separate input processing of this model.

In addition to evaluating chord quality and harmony, an assessment experiment of model structure is also designed. Although the enhanced Transformer-XL model is selected for Pop Music Transformer, which adds fragment-level loop and relative position coding, considering that the transformer in Music Transformer already has relative position coding and the computational space complexity is lower, in order to optimize the computational time and space efficiency, the model structure improvement in this article is based on Music Transformer. Therefore, in the experiment, apart from the above baseline model, the Performance RNN and Transformer model without relative position representation are also introduced to evaluate the performance of the Music Transformer comprehensively. The experimental results are shown in Table 1 (the numbers in brackets represent the model layers).

Model	*NLL*
Performance RNN (3L)	1.899
LSTM for increasing attention mechanism (3L)	1.917
Transformer (6L)	1.801
Music Transformer (baseline model, 6L)	l.778
Improved model (6L+2self-attention)	1.755

Table 1: NLL values of different models.

The models in the above table are arranged according to the proposed time sequence. From the table, it can be seen that the NLL value of the baseline model is less than the results of the first three models, which shows that the performance of the Music Transformer model proposed later is superior to the previous model and it is very reasonable to use Music Transformer as the baseline model of this article. In addition, the NLL value of the model after adding two self-attention layers is the smallest, which shows that the baseline model is optimized by adding two self-attention layers.

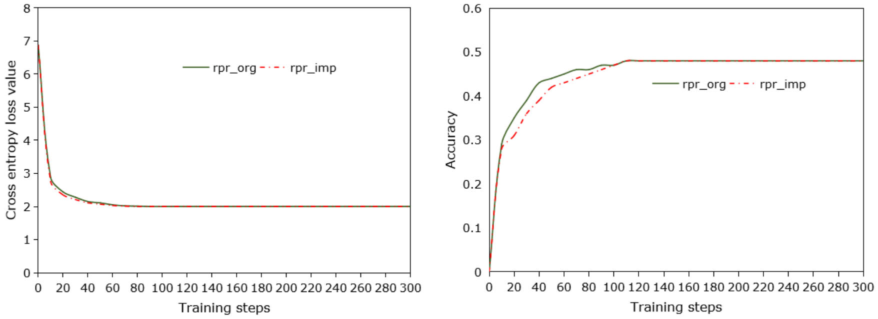

Figure 4: Cross entropy loss value (a) and accuracy (b).

Figure 4 is a visual presentation of the above experimental results, Figure 4(a) is a trend diagram of the loss value of model training, and Figure 4(b) is a corresponding trend diagram of accuracy. It can be seen that the results of the two models are very close, but the loss value of the baseline model is still slightly higher than that of the improved model, so adding two self-attention layers really improves the learning performance of the model.

In the synthesis and optimization of tone quality, deep learning algorithms and CAD technology are used to simulate and enhance the characteristics of tone quality. Because the sound of musical instruments contains rich harmonic information, the harmonic characteristics of music signals are fully utilized in the stage of pitch estimation and timbre simulation. In the pitch estimation stage, only the energy information of the fundamental frequency and its harmonic components is effective for extracting the main melody, while other frequency energies may constitute interference. Figure 5 shows the results of pitch estimation in the presence of fundamental frequency. Through the deep learning model, the fundamental frequency in the music signal can be accurately identified, and the timbre simulation can be carried out accordingly.

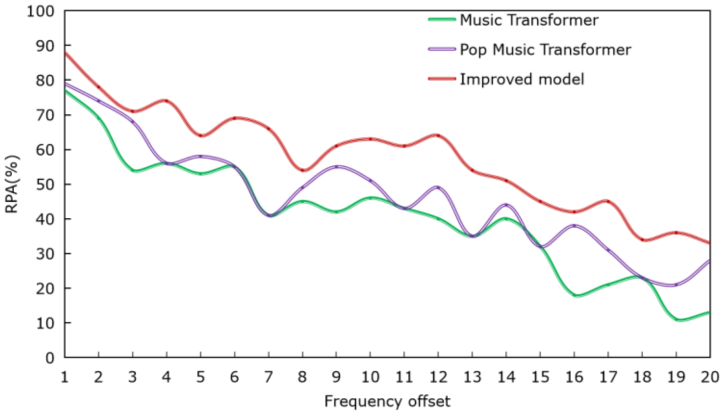

Figure 5: Pitch estimation results in the presence of fundamental frequency.

Figure 5 aims to demonstrate the performance of music converters, popular music converters, and improved models in pitch estimation (fundamental frequency detection). In music signal processing, fundamental frequency (also known as pitch) is the lowest frequency component in a sound signal, which determines the efficiency of automating the music conversion process. Figure 5 contains multiple curves representing the fundamental frequency estimation results of the music converter, popular music converter, and improved model on the same audio signal. By comparing these curves, the accuracy and stability of different models in pitch estimation can be evaluated. It can be seen that the improved model has the best efficiency in process automation. This indicates that CAD technology can be used to simulate the physical structure of musical instruments and deep learning models can be used to learn and imitate specific features of musical instrument sound. In this study, CAD technology is used to simulate the physical structure of musical instruments and a deep learning algorithm is combined to simulate the timbre of musical instruments. By training the deep learning model, we can learn the characteristics of musical instrument sounds and produce timbre similar to real musical instruments when synthesizing.

The creation of multi-tone music signals stems from the overlaying of sound wave patterns generated by diverse musical instruments during recording. One notable challenge lies in the disentanglement of the spectrum originating from various sound sources, aiming to assign them to their respective musical notes. This study tries to solve this problem by using separation technology in deep learning algorithms, such as source separation or sound source localization. Figure 6 shows the comparison of recall rates when using different classifiers for spectrum separation.

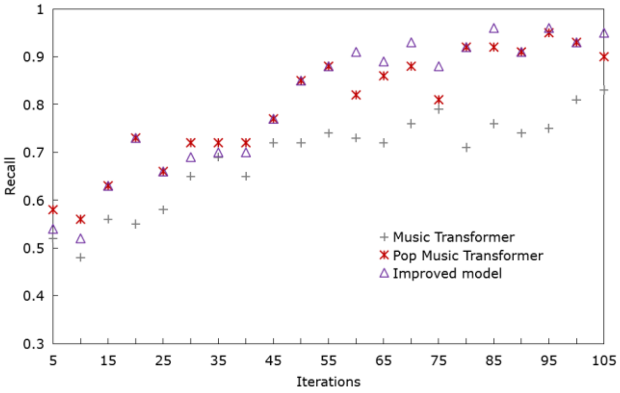

Figure 6: Recall of different classifiers on spectrum separation.

Figure 6 shows the comparison results of recall rates in music signal spectrum separation tasks using different models (including music converters, pop music converters, and an improved model). Figure 6 contains three curves representing the recall performance of the music converter, popular music converter, and an improved model in the spectrum separation task. These curves show the trend of changes in recall rates for each model as the number of iterations increases. By observing these curves, we can compare the performance of different models in spectrum separation tasks. The improved model curve remained above the other two curves throughout the entire iteration process, indicating that the model outperformed the other two models in spectrum separation tasks. By optimizing the structure and parameters of the deep learning model, the accuracy and efficiency of spectrum separation are improved.

Figure 7 illustrates the accuracy of the initial pitch achieved by this method across diverse databases, indicating its consistency and reliability in achieving high accuracy across different data sources. Figure 7 shows the accuracy of different models (including music converters, pop music converters, and an improved model) in estimating initial pitch (or fundamental frequency) on different music databases. Accuracy is an indicator that measures the degree of closeness between the estimated pitch of a model and the actual pitch. The horizontal axis represents different music databases or datasets. Due to the diversity and complexity of music data, the performance of models may vary on different datasets. Therefore, evaluating the performance of the model on different datasets can provide a more comprehensive understanding of its performance. The vertical axis represents the accuracy of pitch estimation.

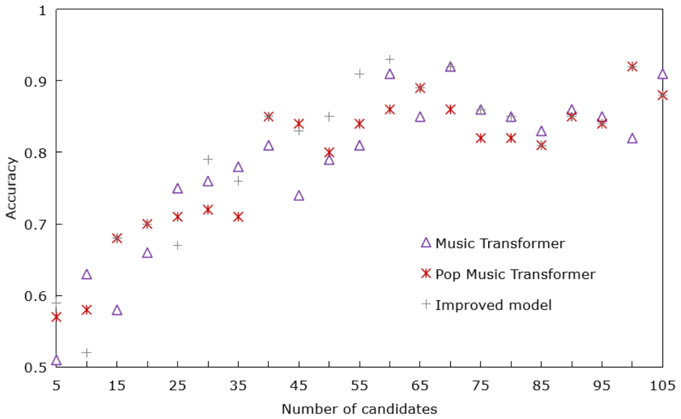

Figure 7: Original pitch accuracy of this method on different databases.

This method quantifies the difference between the model's estimated pitch and the actual pitch. The higher the value, the higher the accuracy of the model. The improved model has the highest accuracy. By adjusting the learning rate, a balance point is found in the study, which ensures the system's stability and improves the training efficiency. This method can accurately estimate pitch, simulate timbre, and effectively separate different instrument sounds in multi-tone music signals.

5 CONCLUSIONS

This study has made breakthrough progress in the field of instrument sound synthesis and optimization, thanks to the clever combination of deep learning algorithms and CAD technology. In the experiment, we continuously optimized the architecture and parameters of the deep learning model, successfully achieving precise separation of different sound source spectra and converting them into corresponding notes. In addition, the use of CAD technology to simulate the physical structure of musical instruments provides valuable data support for deep learning models, significantly improving the quality of sound synthesis. We have delved into the harmonic characteristics of music signals and accurately captured and simulated the pitch and timbre of real instruments through carefully designed deep-learning models. It also tested the accuracy of the original pitch on multiple databases and delved into the impact of learning rate on system performance. And found an ideal balance point, which not only ensures the stable operation of the system but also improves training efficiency. Meanwhile, this study also provides valuable experience and inspiration for the application of deep learning algorithms and CAD technology in other fields. The application of this technology not only improves the accuracy and efficiency of spectrum separation but also opens the door to music innovation for us. It not only provides a new sound synthesis and optimization tool for music producers but also provides rich teaching resources and means for music educators.

Jinxing Mu, https://orcid.org/0009-0003-8047-5945
Jie Tian, https://orcid.org/0009-0004-6375-0332

REFERENCE

[1] Cai, X.; Xi, M.; Jia, S.; Xu, X.; Wu, Y.; Sun, H.: An Automatic Music-Driven Folk Dance Movements Generation Method Based on Sequence-To-Sequence Network, International

Journal of Pattern Recognition and Artificial Intelligence, 37(05), 2023, 2358003. https://doi.org/10.1142/S021800142358003X

[2] Çamcı, A.; Hamilton, R.: Audio-first VR: New perspectives on musical experiences in virtual environments, Journal of New Music Research, 49(1), 2020, 1-7. https://doi.org/10.1080/09298215.2019.1707234

[3] Dong, Y.; Yang, X.; Zhao, X.: Bidirectional convolutional recurrent sparse network (bcrsn): an efficient model for music emotion recognition, IEEE Transactions on Multimedia, 21(12), 2019, 3150-3163. https://doi.org/10.1109/TMM.2019.2918739

[4] Ge, M.; Tian, Y.; Ge, Y.: Optimization of the computer-aided design system for music automatic classification based on feature analysis, Computer-Aided Design and Applications, 19(S3), 2021, 153-163. https://doi.org/10.14733/cadaps.2022.S3.153-163

[5] Georges, P.; Seckin, A.: Music information visualization and classical composers discovery: an application of network graphs, multidimensional scaling, and support vector machines, Scientometrics, 127(5), 2022, 2277-2311. https://doi.org/10.1007/s11192-022-04331-8

[6] Ghisi, D.; Cella, C.-E.: A Framework for modifying orchestral qualities in computer-aided orchestration, Computer Music Journal, 45(4), 2021, 57-72. https://doi.org/10.1162/comj_a_00621

[7] Han, J.: Research on layout optimisation of human-computer interaction interface of electronic music products based on ERP technology, International Journal of Product Development, 27(1-2), 2023, 126-139. https://doi.org/10.1504/IJPD.2023.129315

[8] He, N.; Ferguson, S.: Music emotion recognition based on segment-level two-stage learning, International Journal of Multimedia Information Retrieval, 11(3), 2022, 383-394. https://doi.org/10.1007/s13735-022-00230-z

[9] Huang, Z.; Jia, X.; Guo, Y.: State-of-the-art model for music object recognition with deep learning, Applied Sciences, 9(13), 2019, 2645. https://doi.org/10.3390/app9132645

[10] Jiang, T.: Digital media application technology of mobile terminals based on edge computing and virtual reality, Mobile Information Systems, 2021(14), 2021, 1-10. https://doi.org/10.1155/2021/3940693

[11] Klein, K.; Melnyk, V.; Voelckner, F.: Effects of background music on evaluations of visual images, Psychology & Marketing, 38(12), 2021, 2240-2246. https://doi.org/10.1002/mar.21588

[12] Lattner, S.; Nistal, J.: Stochastic restoration of heavily compressed musical audio using generative adversarial networks, Electronics, 10(11), 2021, 1349. https://doi.org/10.3390/electronics10111349

[13] Liang, Y.; Willemsen, M.-C.: Promoting music exploration through personalized nudging in a genre exploration recommender, International Journal of Human–Computer Interaction, 39(7), 2023, 1495-1518. https://doi.org/10.1080/10447318.2022.2108060

[14] Liao, N.-J.: Research on intelligent interactive music information based on visualization technology, Journal of Intelligent Systems, 31(1), 2022, 289-297. https://doi.org/10.1515/jisys-2022-0016

[15] Liao, Y.; Gui, Z.: An intelligent sparse feature extraction approach for music data component recognition and analysis of hybrid instruments, Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2023(5), 2023, 45. https://doi.org/10.3233/JIFS-231290

[16] Lopes, A.-M.; Tenreiro, M.-J.-A.: On the complexity analysis and visualization of musical information, Entropy, 21(7), 2019, 669. https://doi.org/10.3390/e21070669

[17] Pei, Z.; Wang, Y.: Analysis of computer-aided teaching management system for music appreciation course based on network resources, Computer-Aided Design and Applications, 19(S1), 2021, 1-11. https://doi.org/10.14733/cadaps.2022.S1.1-11

[18] Xiang, Z.; Dong, X.; Li, Y.: Bimodal emotion recognition model for minnan songs, Information (Switzerland), 11(3), 2020,145. https://doi.org/10.3390/info11030145