International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 5671-5684

https://doi.org/10.70517/ijhsa463489

A Study on Multi-Level Accompaniment Effect Generation Based on Timing Data Modeling in Piano Art Instruction

Shuang Du[1,*]
[1] School of Music and Dance, Weifang University, Weifang, Shandong, 261061, China
Corresponding authors: (e-mail: WFXYYWDS@163.com).

Abstract Piano as a popular keyboard instrument is not only a solo instrument but also an important accompaniment instrument. This study explores a multi-level accompaniment effect generation method based on temporal data modeling in piano art instruction. The time-frequency transformation of piano audio by constant Q-transform and short-time Fourier transform realizes the timing data modeling, and builds the accompaniment generation model based on the codec structure to solve the problem of generating the accompaniment tracks based on the main melody and maintaining the melodic harmony among the accompaniment tracks. The study adopts the Lookback mechanism to encode the main melody information, and at the same time utilizes the attention mechanism to realize the coordinated representation of inter-track information. The experimental results show that compared with the MuseGAN and MMM models, the model in this paper achieves a coverage of 0.917 on the note length distribution, which is about 20.0% higher than that of MuseGAN, and a coverage of 0.945 on the pitch distribution, which is about 127.2% higher than that of MMM; In the inter-track distance index, the TD value of piano and guitar is reduced to 0.632, which is much lower than that of MMM's 1.387. The study proves that the model can effectively improve the inter-track harmony while maintaining the quality within the tracks, which is of great significance for the theoretical research and practical application of piano accompaniment.

Index Terms piano accompaniment, temporal data modeling, constant Q transform, multilevel accompaniment, codec structure, attention mechanism

I. Introduction

As a widely popular modern keyboard instrument, the piano is loved and accepted by more and more people, not only as a solo instrument, but also as an important accompaniment instrument [1]. Since the piano's accompaniment function has been explored, many musicians have begun to devote themselves to piano melody creation, and consequently, the status of piano writing accompaniment has been elevated, and the term “piano accompaniment” has gradually come into the public's field of vision [2]. As the name suggests, piano accompaniment refers to the art of expressing the melodic mood and emotion on the keyboard in a short period of time without preparation by the piano user [3], [4]. Although piano accompaniment is a musical art, it contains a variety of factors such as piano playing techniques, accompaniment patterns, harmonic layout, and chord structure [5]. Thus, for learners, the cultivation of piano accompaniment ability is a long-term project that requires professional theoretical guidance and practical training [6], [7]. At present, more and more music enthusiasts are devoted to the research of piano accompaniment principle and practice, attempting to find a more scientific and effective way to cultivate piano accompaniment ability.

At present, domestic and foreign scholars' research on piano accompaniment is mostly focused on the theoretical level, and there are relatively few studies on the practical aspects of piano accompaniment, and the guiding theory of the practical aspects is not rich enough [8]. Therefore, the study of piano accompaniment practice will be the development trend of the discipline for quite a long time in the future [9]. The publication of this book follows this trend and has the effect of targeting the basic guidance of learning methodology, which both broadens the research horizon of piano accompaniment and provides methodological support [10], [11]. In the research on the art and teaching of piano accompaniment, the book can help other piano art enthusiasts and learners to overcome the misunderstanding and action bias, and continue to play a theoretical and practical guiding role, contributing to the further deepening and promotion of the research on the art and teaching of piano accompaniment [12]-[14].

As a widely popular modern keyboard instrument, the piano is loved and accepted by more and more people, not only as a solo instrument, but also as an important accompaniment instrument. Since the piano's accompaniment function has been explored, many musicians have started to devote themselves to piano melody creation, and consequently, the status of piano writing accompaniment has been elevated, and the term “piano accompaniment” has gradually entered the public's field of vision. Piano accompaniment refers to the art of expressing melodic moods and emotions on the keyboard in a short period of time without preparation by the piano user. Although piano

accompaniment is a musical art, it contains a variety of factors such as piano playing techniques, accompaniment patterns, harmonic layouts and chord structures. Therefore, for learners, the cultivation of piano accompaniment ability is a long-term project, which requires professional theoretical guidance and practical training. At present, more and more music lovers are devoted to the research of piano accompaniment principle and practice, trying to find a more scientific and effective way to cultivate piano accompaniment ability. At present, scholars at home and abroad focus on the research of piano accompaniment at the theoretical level, and there are relatively few researches on the practical aspects of piano accompaniment, and the guiding theory of the practical aspects is not rich enough. Therefore, in the future for quite a long time, the study of piano accompaniment practice will be the development trend of this discipline. In the research aimed at the art and teaching of piano accompaniment, the relevant research can help piano art enthusiasts and learners to overcome the misunderstanding and action bias, continuously play a guiding role in theory and practice, and contribute to the further deepening and promotion of the research on the art and teaching of piano accompaniment. With the development of artificial intelligence technology, it has become possible to utilize computer-assisted piano accompaniment generation, which not only provides diversified accompaniment choices, but also provides piano learners with more intuitive learning materials, thus promoting the cultivation of piano accompaniment practice ability.

This study focuses on the field of piano art instruction, and optimizes the generation of multi-level piano accompaniment effects through time-series data modeling techniques. The study first applies the constant Q transform and the short-time Fourier transform to the time-frequency transformation of the piano audio to achieve more accurate time-sequence data modeling. On this basis, a codec-based accompaniment generation model is designed, which solves two core problems: how to generate each accompaniment track based on the main melody, and how to maintain the melodic harmony among the accompaniment tracks. For the information representation of the main melody, the Lookback mechanism is used, and the key features of the main melody are extracted by the attention mechanism; for the information representation of the multi-tracks, the parallel fully-connected layer is used to generate the backing tracks synchronously, to ensure that the information is shared among the tracks. The model structure adopts a combination of an encoder with added attention mechanism and a multi-track decoder, which realizes multi-level and high-quality piano accompaniment generation through the synergy of attention vectors, implicit layer states and multiple parallel fully-connected layers. This study will explore the theory and practice in depth, with a view to providing new methods and ideas for piano accompaniment teaching and practice.

II. Timing data modeling of piano audio

In order to improve the multilevel effect of the generated piano accompaniment, before using the model to generate the accompaniment, this paper applies the constant Q-transform and the short-time Fourier transform to perform the time-frequency transform on the acquired piano audio, so as to realize the temporal data modeling.

II. A. Constant Q Transform (CQT)

The constant Q transform (CQT) [15] is a time-frequency analysis technique widely used in music signal processing and acoustic research. It decomposes the frequency of a signal by a set of filters, which are characterized by an exponentially regular distribution of the center frequency and different filter bandwidths, but the ratio of the center frequency to the bandwidth is a constant Q. This means that in CQT, the ratio of the center frequency of each filter to its bandwidth is fixed, so that the bandwidth of the filters increases as the center frequency increases in different frequency ranges to keep the value of Q constant.

The CQT's spectral cross-axis frequency uses a logarithmic scale based on a base of 2 instead of a linear scale, which matches the distribution of musical scales and allows the CQT to better capture subtle frequency variations in the audio signal. Since the frequency distribution of music is usually nonlinear, CQT has significant advantages in music signal processing.

For a constant Q filter, the ratio of the center frequency to the bandwidth is a fixed value and can be expressed by the following equation:

$$Q = \frac{f_c}{\Delta f} \tag{1}$$

where Q is the value of the constant Q, f_c is the center frequency of the filter, and Δf is the bandwidth of the filter.

The bandwidth of the filter and the spacing between neighboring frequencies are adjusted to ensure that the frequency resolution in different frequency ranges can adapt to changes in signal characteristics. For low-frequency waveforms, CQT will use a narrower filter bandwidth to enhance the resolution of notes with small frequency intervals. For high frequency waveforms, on the other hand, the CQT will use a wider filter bandwidth to enhance the temporal resolution for rapidly changing overtones.

By definition, the frequency bandwidth δ_f at frequency f, also known as frequency resolution, indicates the filter bandwidth at that frequency. In CQT, the bandwidth of the filter varies with frequency to ensure that the frequency resolution adapts to changes in signal characteristics over different frequency ranges.

Assuming that the lowest tone to be processed is $f_{\min}$, the frequency of the k th frequency component, f_k, can be expressed by the following equation:

$$f_k = f_{\min} \cdot 2^{k/b} \tag{2}$$

where b denotes the number of spectral lines contained within each octave.

In CQT, the frequency resolution δ_f can be expressed in terms of the bandwidth of the filter. For the frequency bandwidth that is at frequency f_k, it is usually defined as:

$$\delta_{f_k} = Q \cdot f_k \tag{3}$$

where Q is a constant Q value that represents the ratio of the center frequency f_k of the filter to the bandwidth δ_{f_k}. Then it is known from the above equation:

$$Q = \frac{f}{\delta_f} = \frac{1}{2^{1/b} - 1} \tag{4}$$

Therefore, the value of Q is related to b.

According to the given conditions, the window length N_k with frequency can be calculated as follows:

$$N_k = \left\lceil Q \frac{f_s}{f_k} \right\rceil, k = 0,1,\cdots,K-1 \tag{5}$$

where $\lceil x \rceil$ denotes the upward rounding function, f_k is the frequency of the k th semitone, f_s is the sampling frequency, and K is the total number of semitones.

To summarize, so in CQT, the k th semitone frequency component of the n th frame can be expressed as:

$$X^{CQT}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x(n) w N_k(n) e^{-j\frac{2\pi Q}{N_k} n} \tag{6}$$

where $X^{CQT}(k)$ denotes the frequency component of the k th semitone, $x(n)$ is the sampled value of the input signal in the time domain, and $wN_k(n)$ is a window function of length N_k.

II. B.Short-Time Fourier Transform (STFT)

The Fourier transform is an important tool for signal processing, mainly used to convert signals from the time domain to the frequency domain, but it fails to provide the localized characteristics of the signal in the time domain. In order to solve this problem, Short Time Fourier Transform (STFT) is proposed [16].STFT divides the signal into multiple time segments and applies a window function weighting in each time segment and then performs a Fourier Transform to locally analyze the signal in the time-frequency domain.

The process of STFT is to multiply the signal by a time-limited window function $h(t)$ before the signal is Fourier transformed. This window function serves to limit the time horizon of the signal in the time domain and assumes that the signal is smooth within the analysis window. The signal is then analyzed segment by segment by shifting the position of the window function $h(t)$ on the time axis to obtain a set of localized spectral information.

The mathematical expression for the STFT is:

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau) \cdot h(\tau - t) \cdot e^{-j\omega\tau} d\tau \tag{7}$$

where $X(t,\omega)$ denotes the spectral component at time t and frequency ω and $x(\tau)$ denotes the original signal. The $h(\tau - t)$ is a window function, which is usually 1 at the moment t and decays to zero at other moments, used to limit the range of the signal in time. This step helps to minimize spectral leakage and ensures a smooth transition of the signal to zero at the window boundary. The window function is generally chosen as a Hanning window, a

Hamming window, etc. In particular, when the window function is taken as $h(t) \equiv 1$, the STFT is equivalent to a conventional Fourier transform.

II. C.MIDI to CQT spectrograms

For the input of piano notated music in MIDI format, at this stage of the paper an encoder decoder architecture (Midiff) is designed as shown in Figure 1. The first encoder consists of a self-Attend layer and an MLP layer, which is responsible for receiving a series of symbolic note events, which can be note events containing any number of instruments. The second encoder, on the other hand, also consists of a self-Attend layer and an MLP layer, which can optionally use the early part of the Meier spectrogram as contextual information. This information is passed to the decoder, whose task is responsible for generating a CQT spectrogram corresponding to the input note sequence. In this paper, the diffusion model is trained as a decoder.

Diffusion model is a probabilistic generative model that iteratively generates data from noise by reversing the Gaussian diffusion process. The model consists of two main parts, the forward noise addition process and the reverse denoising process. In the forward process, the input signal x is converted to noise $\varepsilon \sim N(0, I)$, which occurs at diffusion time step $t \in [0,1]$. The resulting noise figure X_t is thus given by the following equation:

$$X_t = \alpha_t X + \sigma_t \grave{o} \tag{8}$$

Where $\alpha_t \in [0,1]$ and $\sigma_t \in [0,1]$ are the parameters in the noise table used to mix the original signal and the noise in diffusion time. In this work, the decoder $\grave{o}_\theta$ is trained to predict the additive noise given noisy data. To achieve this, this paper by minimizing an objective loss function of the form:

$$L_{Midiff} = \underset{X,c,\grave{o},t}{E} w_t \left\| \grave{o}_\theta \left(X_t, c, t \right) - \grave{o} \right\|_1^1 \tag{9}$$

where w_t is a set of loss weights to weight the losses for different diffusion time steps. The c is additional condition information for the decoder. These weights w_t, the parameter α_t, and the time step t are hyperparameters used to selectively emphasize particular steps in the backward diffusion process.

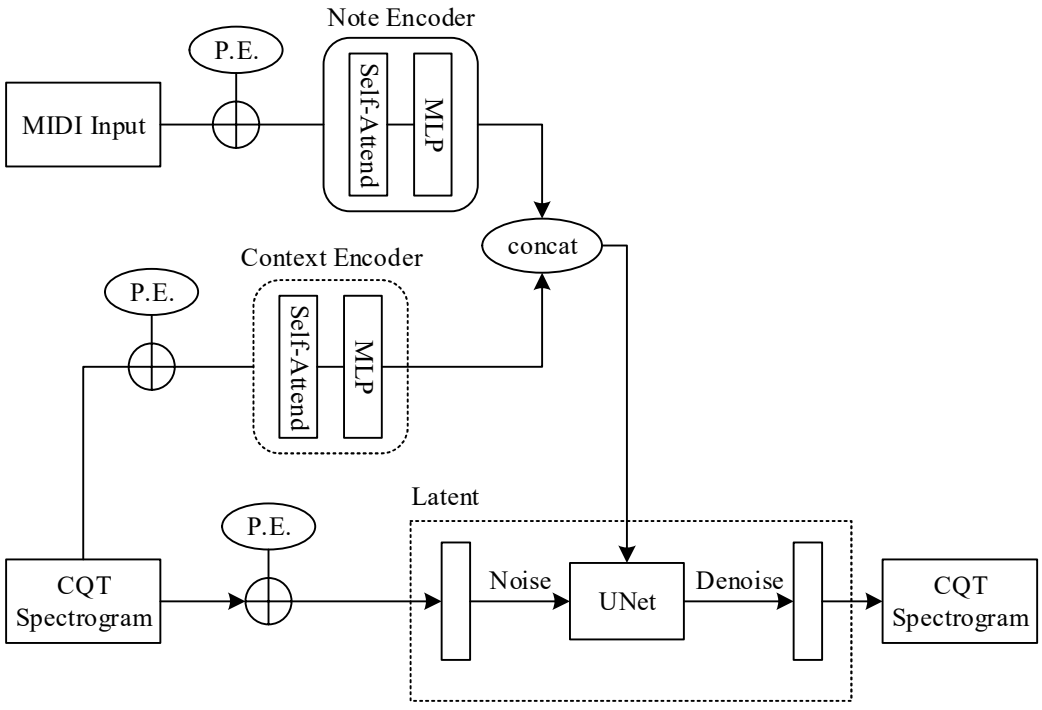


Figure 1: Midiff Architecture

During the sampling process, this paper follows the reverse diffusion process. Starting with independent Gaussian noise for each frame and frequency bin, the noise estimation is used iteratively to gradually reduce the noise content

and generate a new CQT spectrogram. During the inference process, this paper scales the model output to the expected range of CQT spectrograms.

In this work, this paper uses a one-dimensional U-Net architecture at the decoder stage as shown in Fig. 2. Where (R) represents the residual 1D convolution unit, which is used to learn the features efficiently. (M) represents the modulation unit, which is used to change the channel for a given feature at different diffuse noise levels. (I) represents the injection item, which connects the external channel to the current depth in order to deliver the impact. (A) represents attention items, used to share contextual information. (C) represents cross attention items for learning text embedding conditions. In this paper, each frequency is treated as a different channel, allowing U-Net to be successfully applied to CQT spectrograms.

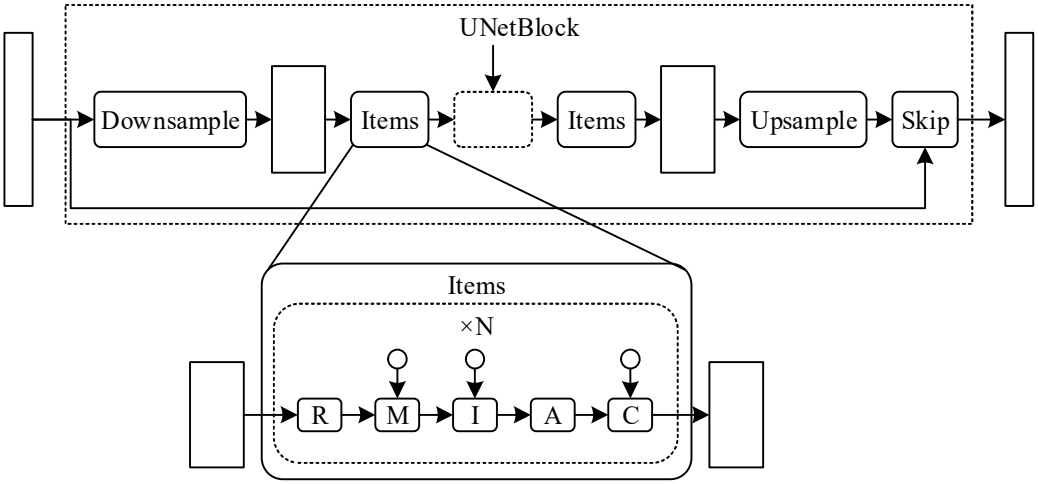


Figure 2: One-dimensional U-Net architecture

II. D. Time-frequency analysis of piano audio data

II. D. 1) Audio Acquisition

The core idea of this research is to convert the audio into a spectrogram to obtain the timing data of the piano audio, so as to improve the multilevel accompaniment generation effect of the piano accompaniment generation model. Therefore, when selecting data for the training set, the audio timings should preferably be of the same size and not too long. Because the online data set that meets the requirements is too small, the sample is not enough, so the first step of this experiment is to segment the audio. Take Beethoven's Moonlight Sonata as an example, each audio clip obtained from segmentation is 6 seconds long.

After obtaining enough small segments of audio, time domain and frequency domain analysis is performed. One small fragment is selected as an example, and the time domain plot based on the first 6-second fragment of Beethoven's Moonlight Sonata is shown in Figure 3. Its horizontal coordinate is time and vertical coordinate is amplitude.

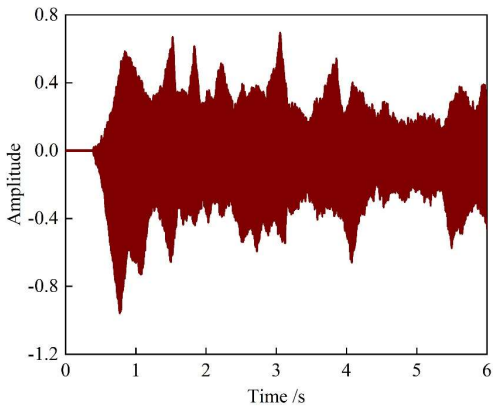


Figure 3: Time-domain graph of piano audio clips

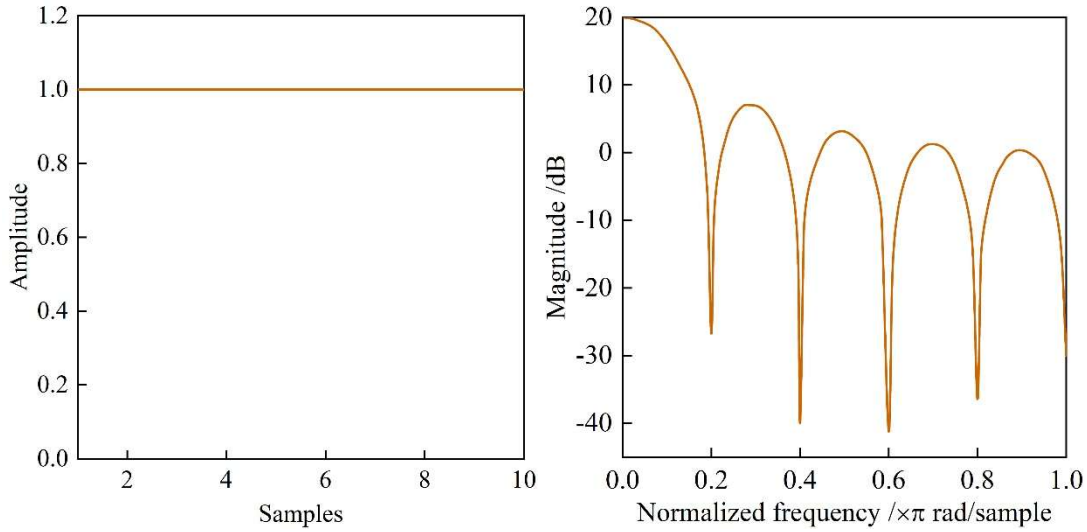
II. D. 2) Spectrograms of short-time Fourier transforms

According to the conjecture before the experiment, the short-time Fourier transform should not be as effective as the constant Q transform in processing the piano music signal. Considering that the short-time Fourier transform is one of the most commonly used algorithms, it is used as a control experiment. Compared to the Fourier transform, the short-time Fourier transform adds a window function. The window function consists of many different types, the common ones are Hanning window, Kaiser window, rectangular window and so on.

(1) Selection of window function

1) Rectangular window

Rectangular window is the most frequently used window function, but also the default use of the window function. Rectangular window is suitable for the scenario that only needs the main flap frequency, and the amplitude accuracy has little effect. After a waveform is added with a rectangular window, the time domain and frequency domain diagrams obtained are shown in Fig. 4, where (a) and (b) are the time domain and frequency domain diagrams, respectively, and the same afterward.

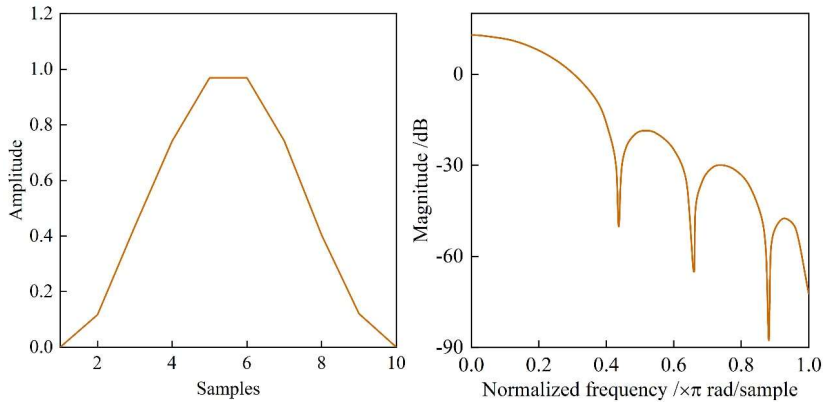


(a) Time domain (b) Frequency domain

Figure 4: Time-domain and frequency-domain graphs processed by rectangular Windows

2) Hanning Window

The Hanning window is applicable to non-periodic continuous functions and is generally considered to be a special case of the ascending cosine window. The time-domain and frequency-domain plots obtained after adding a Hanning window to a certain waveform are shown in Figure 5.

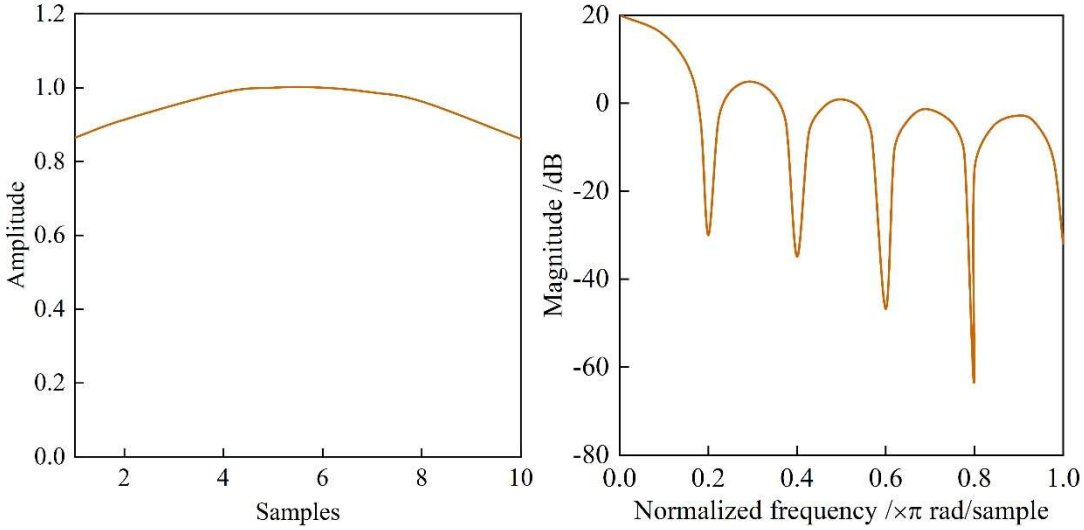


(a) Time domain (b) Frequency domain

Figure 5: The time domain and frequency domain processed by the Hanning window

3) Kaiser Window

The Kaiser window is strictly a group of functions containing an adjustable parameter, α, which adjusts the width of the main and side flaps by adjusting the size of the α. The Kaiser window is a group of functions that contains an adjustable the number, the parameter that adjusts the window. After adding the Kaiser window to a certain waveform and setting the parameter $\alpha=0.8$, the time domain and frequency domain plots obtained are shown in Fig. 6.



(a) Time domain (b) Frequency domain

Figure 6: The time domain and frequency domain processed by the Kaiser window

After comparing several kinds of window functions, considering that the piano music has many frequency components and the spectrum is complex, this experiment adopts the Hanning window, and sets the length size of the window to 512.

(2) Drawing short-time Fourier transform spectrograms

Take Beethoven's “Moonlight Sonata” as an example again, take its first 6 seconds of the clip, respectively, to get the time-frequency spectrum of the spectrogram shown in Figure 7. The horizontal coordinate in the graph is time, the vertical coordinate is frequency, and different colors represent different energies, which can be regarded as different amplitudes. From the speech spectrogram, it is not possible to directly analyze the meaning of the two graphs, and it is not possible to see what connection there is between the speech spectrogram and the timbre, let alone manually extracting timbre-related features from the speech spectrogram. However, it can be clearly seen that the amplitude of the spectrogram varies with time, which indicates that the purpose of time series data modeling can be achieved by transforming the spectrogram.

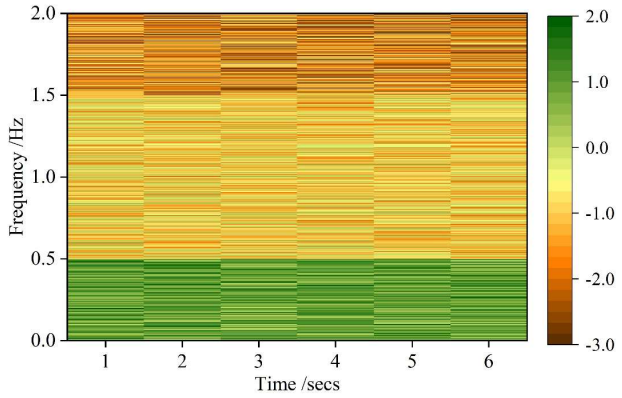


Figure 7: Spectrogram of the time-spectrum

II. D. 3) Spectrograms of constant Q-transforms

The constant Q transform has better resolution for frequency components with lower pitches and performs better when dealing with some bass instruments. The pitch and frequency comparison is shown in Table 1, C2 and C#2 differ by a semitone, and their frequencies differ by about 5 Hz, when the pitch is raised by 3 octaves, the same difference is a semitone, and the frequency difference between C5 and C#5 is about 32 Hz. The resolution of low-frequency pitch is very unsatisfactory when the scale of the coordinate axis is linear. The constant Q-transform scale is not linear, but rather a non-linear scale with a logarithmic base, and it will perform better in processing piano audio data.

Table 1: Comparison of pitch and frequency

C2	C#2	C5	C#5
66.5Hz	71.3Hz	535.6Hz	567.5Hz

In plotting the constant Q-transformed speech spectrogram, the opening 6 seconds of Beethoven's Moonlight Sonata were continued to be used in order to control the variables. The results show that the spectrogram obtained by the constant Q-transform method has an exponentially increasing vertical coordinate and is not linear.

III. Multi-level accompaniment generation model for piano based on codec structure

Aiming at the application scenario of piano art instruction, this paper proposes an accompaniment generation model based on the codec structure on the basis of the piano audio timing data modeling, which mainly solves two problems, namely, how to let each accompaniment track be generated based on the main melody, and how to maintain the melodic harmony between each accompaniment track.

III. A. Information representation of the main theme

For the generated main theme music, this section uses the Lookback mechanism to encode a total of 140 MIDI events. In addition, since the encoder and decoder need to use additional symbols <bos> and <eos> to denote the start and end of the sequence, the dictionary size of the whole model is 142. The application of the Attention mechanism requires the encoder to record the hidden layer states at each time step when encoding the main melody, so that the decoder can compute the attention at different moments by using these hidden layer states to derive the attention vector at different moments. Therefore, after adding the <eos> symbol to the end of the main theme, a sequence of length n can be encoded by the encoder to produce n hidden layer vectors. The main melody vector encoding process is shown in Fig. 8.

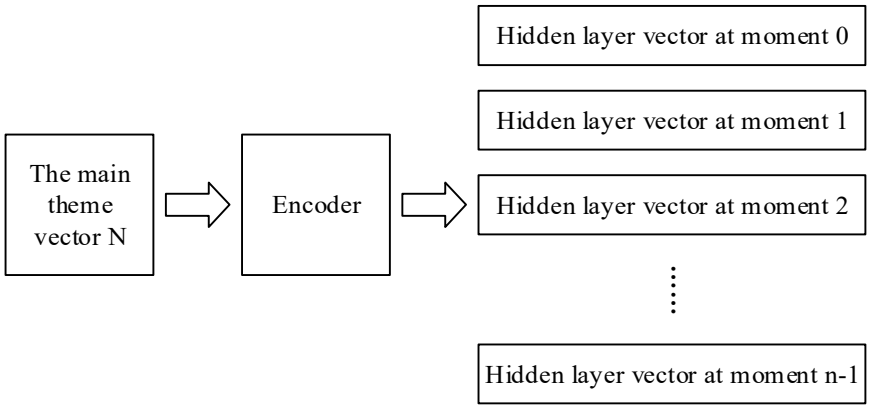


Figure 8: The encoding process of the main melody vector

III. B. Inter-track information representation for multi-tracks

In order to make music more appealing and expressive without being monotonous, it is necessary to have multiple instruments accompanying the performance. Since the information between multiple tracks is interconnected, if each track is modeled and generated independently, the tracks may only collaborate with the main melody, but lack the connection between them, giving the accompaniment a disorganized feel as a whole.

Therefore, it is important to create models that allow each track to capture the information of the main melody while also having the ability to access the information of the other backing tracks. The backing tracks do not prioritize each other, so they are generated synchronously in the decoder. This requires that for a single accompaniment

track, in addition to the necessary information about the main theme and the generated events of its own track, it must also contain information about the generated events of other accompaniment tracks when generating the note events of the current time step.

The model designed in this paper takes as input all the note events of the previous time step of each accompaniment track in the decoder part. After the decoder implicit layer obtains the current implicit layer state from the main melody information vector, the previous implicit layer state, and the input vector, it connects multiple fully-connected layers in order to generate the note events for the instruments of each track at the current moment. Since the information obtained by each fully connected layer contains all the information of the main melody and other accompaniment tracks, the model models the accompaniment tracks as a whole, so that the generation of note events for each track can take into account the requirements of the main melody and the accompaniment parts to work harmonically together. The inputs and outputs of the decoder in a single step are shown in Figure 9.

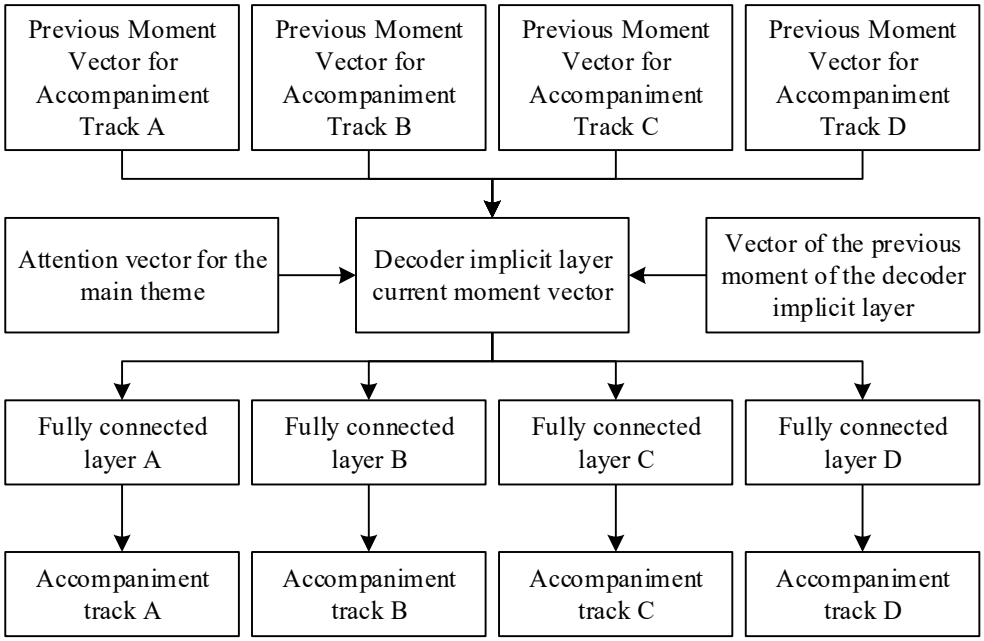


Figure 9: Illustration of input and output of the decoder

III. C. Overall model structure

The accompaniment generation model proposed in this paper contains two main parts, the main melody encoder part with added attention mechanism and the multi-track decoder part. The encoder is mainly responsible for encoding the main melody and representing the information of the main melody in the form of multiple vectors. The decoder part is mainly responsible for generating each accompaniment track.

The single-step process of model generation consists of three main stages. Firstly, based on the implicit layer state of the decoder at the previous moment and the implicit layer state at each moment of the encoder stage, the attention vector of the main melody in the current state is computed by the Attention mechanism. After that, the attention vector is used to calculate the implied layer state of the decoder at the current moment by combining the implied layer state of the decoder at the previous moment and the merged vectors of the outputs of the accompaniment tracks at the previous moment. Finally, this implied layer vector is used to calculate the current moment's output value of each accompaniment track through multiple parallel fully connected vectors. The whole process can be represented as:

$$\alpha_t' = Soft\max\left(v \cdot \tanh\left(W_{s\alpha} s_t' + W_{h\alpha} h_{t-1}\right)\right) \tag{10}$$

$$c_t = \sum_{t'=0}^{T} \alpha_t' s_t' \tag{11}$$

$$i_{ik} = \sigma\left(W_{hik} h_t + b_{ik}\right) \tag{12}$$

$$I_t = Concat\left(i_{t0}, i_{t1}, \cdots, i_{tK}\right) = \left(i_{t00}, i_{t01}, \cdots, i_{tke}, \cdots, i_{tKE-1}, i_{tKE}\right) \tag{13}$$

$$h_t = g\left(I_{t-1}, c_t, h_{t-1}\right) \tag{14}$$

where Eqs. (10) to (11) represent the computation of the attention mechanism. v, $W_{s\alpha}$, $W_{h\alpha}$ are the model trainable parameters, s_t' denotes the state of each implicit layer generated by the encoder, and h_{t-1} denotes the state of the implicit layer of the decoder at the moment $t-1$. α_t' denotes the attention mask assigned to the encoder's implicit layer state s_t' for the t' moment of the encoding phase, which, since it is a Softmax result, is a value between $[0,1]$ and the sum of all α_t' sums to 1. Multiplying and summing the corresponding α_t' with s_t' yields c_t, which is the attention vector for the main melody computed at moment t. Let the accompaniment have a total of $K+1$ different channels and the total number of instrumental note events is $E+1$. Eq. (12) represents the computation of the note events generated by the k th instrument at the t th moment, where W_{hik} and b_{ik} are the trainable parameters of the model, and i_{tk}, computed by the Sigmoid function, represents the output of the k th instrument at the t th moment. The I_t in Eq. (13) is the merger of the output results of all accompanying instruments at the moment t, which is obtained by splicing $K+1$ vectors of output results obtained in Eq. (12). The implicit layer state h_t of the decoder part is shown in Eq. (14) and receives inputs from three parts, namely the output of the previous moment, the attention vector and the decoder implicit layer vector of the previous moment. The g function is computed in such a way that it is determined by the cell structure of the decoder itself.
The model uses a gated recurrent unit (GRU) as the unit of the recurrent neural network, and its computation of updating the state of the implicit layer, i.e., the g function, can be expressed as:

$$r_t = \sigma\left(W_{Ir} I_{t-1} + W_{hr} h_{t-1} + W_{cr} c_t + b_r\right) \tag{15}$$

$$z_t = \sigma\left(W_{Iz} I_{t-1} + W_{hz} h_{t-1} + W_{cz} c_t + b_z\right) \tag{16}$$

$$\tilde{h}_t = \tanh\left(W_{Ih} I_{t-1} + W_{hh}\left(h_{t-1} \odot r_t\right) + W_{ch} c_t + b_h\right) \tag{17}$$

$$h_t = z_t \odot h_{t-1} + \left(1 - z_t\right) \odot \tilde{h}_t \tag{18}$$

where the $\odot$ symbols in Eqs. (17)~(18) denote the operation of multiplying two homogeneous matrices by their elements.The GRU unit contains the concepts of reset and update gates.In Eq. (15) W_{Ir}, W_{hr}, W_{cr}, and b_r are the model trainable parameters, and the computation result r_t is the output parameter of the reset gate. In Eq. (16) W_{Iz}, W_{hz}, W_{cz} and b_z are model trainable parameters, and the computed result z_t is the update gate output parameter. Since both are obtained after computation via Sigmoid function, they are all values between $[0,1]$. The $\tilde{h}_t$ in Eq. (17) denotes the candidate hidden layer state, which is obtained from the inputs at the current moment, the attention vector at the current moment, and the hidden layer state at the previous moment with the reset gate parameter operation, and the W_{Ih}, W_{hh}, W_{ch}, and b_h are all model trainable parameters. Finally, the new hidden layer state h_t in Eq. (18) is obtained from the hidden layer state at the previous moment and the candidate hidden layer state at the current moment jointly computed with the update gate parameters.

III. D. Analysis of the effect of multilevel accompaniment generation in piano

In order to verify the effect of the proposed model piano accompaniment generation, this paper conducts a comparison experiment between this model and other multi-track music accompaniment models MuseGAN [17] and MMM.

III. D. 1) Evaluation indicators

In this paper, the following quantifiable features are chosen as evaluation metrics for the effectiveness of piano accompaniment generation. The metrics for measuring the quality of melody generation within a track are as follows:

Number of pitches (PC): calculates the number of pitches per bar of the music fragment.
Average pitch interval (PS): calculates the pitch transition between two neighboring notes, one semitone as a unit.
Average Intonation Interval (IOI): average of the time interval between two consecutive tones per measure.
Polyphony Rate (PR): the ratio of the number of time steps in which multiple pitches are located to the total number of time steps, ignored for drum tracks, calculated as follows:

$$PR = \frac{\text{Number of polyphony time steps}}{\text{Total time steps}} \tag{19}$$

The above metrics only measure the quality of note generation within a single track. To measure inter-track harmonization, in this paper, tracks are combined two-by-two as a pair and measured using the inter-track distance metric (TD), which is as small as possible.
Intra-track metrics and inter-track harmonization are measured by examining the distribution of pitches and note lengths in the generated accompaniment dataset. The pitch histogram and note length histogram metrics are first calculated for all generated accompaniments, then the distance of each music clip from the rest of the collection is calculated, a probability distribution function is fitted, and finally the area of coverage is calculated with respect to the distribution of the real dataset, with a larger value being better. x represents any one feature of the scale histogram or note length histogram, $P_{i,j}^{x}$ represents the x metrics on the real dataset, and $\hat{P}_{i,j}^{x}$ is the generated data metrics, which are calculated as follows:

$$D_x = \frac{1}{N_{tracks} \cdot N_{bars}} \sum_{i=1}^{N_{tracks}} \sum_{j=1}^{N_{bars}} OA\left(P_{i,j}^{x}, \hat{P}_{i,j}^{x}\right) \tag{20}$$

Pitch Histogram (PCH): an octave-independent representation of pitch with a dimension of 12. It represents the octave-independent chromatic quantization of the frequency continuum.
Note Length Histogram (NLH): calculates the distribution of the length of each note.

III. D. 2) Experimental results and analysis

In order to measure the model effect, this paper in LPD, Freemidi and GPMD three datasets randomly selected 1000 clips as a test case, respectively, using MuseGAN, MMM and this paper's model, according to the main melody of the piano, to generate the accompaniment of the drums, guitars, basses, and strings, to generate the music in the same format as the input to generate the same 4/4 beat.
In this paper, the Lookback mechanism is used to encode in the data processing process, in order to verify its effectiveness, the experimental process is set up in the model itself to compare the experiments, using Textual to indicate the training effect of this paper's model without the use of Lookback, and Textual + Lookback to indicate the effect of the use of the model.
The results of generating quality measurements of music clips within the tracks are shown in Table 2, and the experimental results are calculated in units of 1 bar, taking the mean values of the four accompaniment tracks of drums, guitar, bass and strings.Ground-Truth represents the indicator results of the original dataset, which is used for reference and comparison. In order to visualize the difference with real music data, the last four rows represent the difference between the experimental data generated by different models and the Ground-Truth data, theoretically the smaller the absolute value means the smaller the gap, and the bolded data is the best result.
Comparing the MuseGAN and MMM models, the model in this paper has a smaller gap with the real dataset Ground-Truth on the metrics of pitch usage, pitch shift, note spacing and polyphony rate within the track, which is closer to the real accompaniment situation. Compared to itself, Textual+Lookback, which uses a new encoding method, can further improve the quality of the generation, which is analyzed because the marking of note onsets preserves the musical material in its most pristine condition and avoids the confusion between continuous notes and long notes.
Harmony between tracks is also an important indicator of the effectiveness of accompaniment generation. The track relationships between the piano and the four tracks were evaluated in the comparison experiments, and the results of the harmony metrics are shown in Table 3, where each instrument is represented by its initial letter, i.e., P, D, G, B, and S for piano, drums, guitars, basses, and strings, respectively. The harmony of the accompaniment generated by the three models is compared by calculating the distance TD between each pair of tracks, and the optimal results are indicated by bolding.
During the experiment, it was found that the Lookback mechanism encoding method has almost no effect on the inter-track distance, so the own comparison results of this paper's model were not set. Among the three models, the results of this paper's model are relatively the best, especially in the performance on piano and guitar and bass, and the difference between the tracks is obviously smaller than the other two models, but there are still different degrees of gaps between the performance of different datasets and real music data.

Table 2: The quality of music segments within the orbit

Data set	Model	PC	PS	IOI	PR
LDP	Ground-Truth	3.652	4.058	2.416	0.445
	MuseGAN	+4.694	+11.859	-0.857	+0.564
	MMM	+0.767	+1.288	+0.285	-0.173
	Textual	-0.482	-1.073	+0.212	-0.156
	Textual+Lookback	-0.465	-0.998	+0.105	-0.158
FreeMidi	Ground-Truth	3.715	4.796	2.562	0.463
	MuseGAN	+4.415	+7.058	-1.008	+0.547
	MMM	+0.989	+1.152	-0.391	+0.205
	Textual	-0.432	-1.143	+0.168	-0.151
	Textual+Lookback	-0.409	-1.094	-0.112	-0.149
GPMD	Ground-Truth	3.715	4.796	2.562	0.463
	MuseGAN	+4.386	+6.632	-0.933	+0.615
	MMM	+0.718	+3.494	-0.158	-0.189
	Textual	-0.656	-1.807	-0.051	-0.157
	Textual+Lookback	-0.445	-1.598	-0.039	-0.090

Table 3: Inter-orbital generation mass

Data set	Model	P-D	P-G	P-B	P-S	B-G	B-S
LDP	Ground-Truth	1.657	0.965	1.624	0.741	1.145	1.001
	MuseGAN	0.928	1.174	1.245	1.178	0.826	1.029
	MMM	1.496	1.387	1.603	1.165	1.739	1.707
	Textual	0.923	0.632	1.052	0.784	0.611	0.679
FreeMidi	Ground-Truth	1.629	1.038	1.535	0.431	1.208	0.516
	MuseGAN	0.747	1.205	1.326	1.294	0.852	1.051
	MMM	1.591	1.292	1.769	1.503	1.567	1.846
	Textual	0.637	0.728	0.965	0.997	0.779	0.768
GPMD	Ground-Truth	1.527	0.786	1.679	0.783	1.022	0.993
	MuseGAN	0.781	1.185	1.321	1.294	0.854	1.039
	MMM	1.465	1.281	1.394	1.117	1.458	1.575
	Textual	0.751	0.729	0.907	0.809	0.625	0.673

Overall, the MuseGAN model track dependency is better than MMM, but the generated single-track effect is somewhat distant from the real dataset, especially on the pitch-to-pitch transitions.The MMM model, although it can generate music clips with higher quality within the tracks, has weaker inter-track connections. The model in this paper, on the other hand, outperforms both in terms of piano accompaniment generation quality and track harmony. The pitch and note length distributions reflect the melodic direction of a piece of music. In order to further quantify the differences between the musical pieces and the real dataset, this paper calculates the distributions of the scale histograms and note length histogram features of all the piano accompaniments generated by the three models, which are converted to probability distribution functions using kernel density estimation and plotted as Figs. The probability distribution functions of the pitch lengths and the probability distribution functions of the pitch lengths of the three models and the real data are shown in Fig. 10 and Fig. 11, respectively.

It can be seen that the model generation effect of MMM has a larger variance in pitch and duration, and the mean value slightly deviates from the real data, while MuseGAN is close to the real music, but the variance is smaller.The use of Lookback mechanism coding method has a certain positive impact on the model generation effect in this paper, and the difference between the two data distributions is not large.

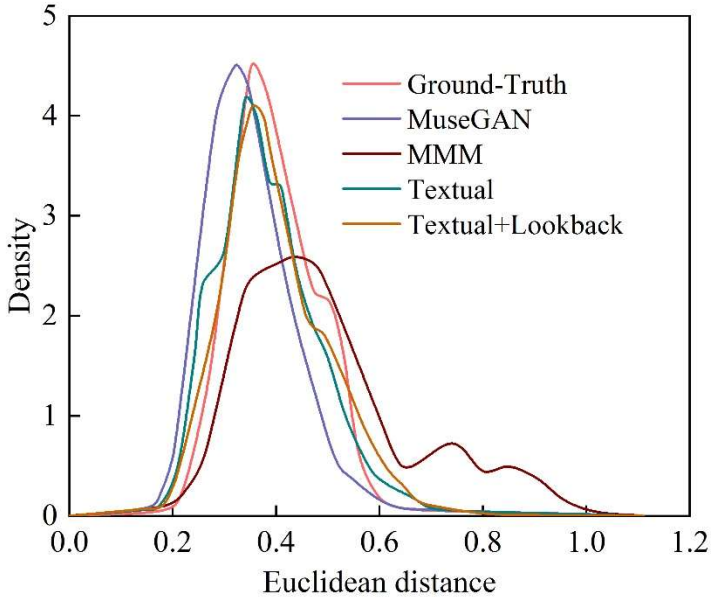


Figure 10: Probability distribution map of note length

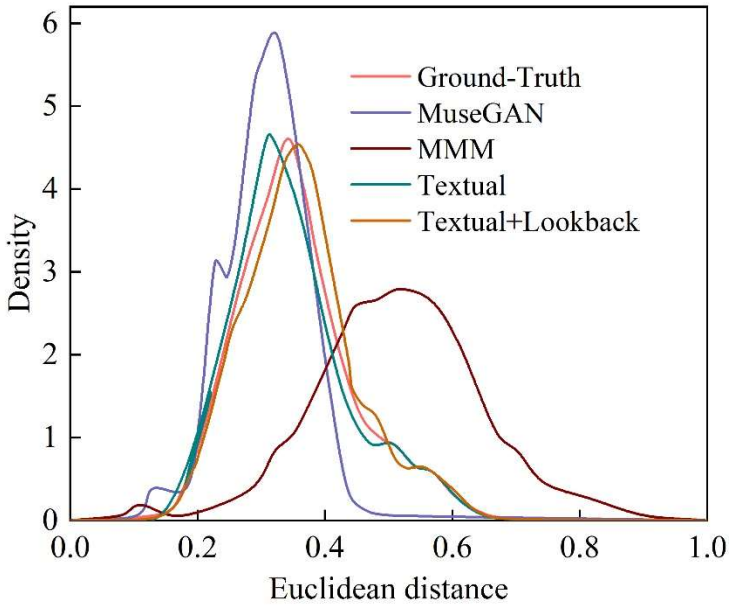


Figure 11: Pitch probability distribution map

In order to observe the gap more intuitively, this paper calculates the overlap area between the data generated by each model and the distribution function of the real dataset. The results of the overlap area calculation for the three models are shown in Table 4, the larger the value the higher the coverage and the closer it is to the real music accompaniment. Same as the observation results of the distribution map, MuseGAN is better than MMM, but inferior to this paper's model, this paper's model obtains the highest values, NLH and PCH are 0.894 and 0.941 respectively, which are closer to the real data, Lookback mechanism coding method can further improve the similarity to a certain extent, after using Lookback mechanism coding the model of this paper's NLH and PCH are improved by 0.023 and 0.004 respectively.

Table 4: OA Distances on NLH and PCH (Unit: percentage)

	MuseGAN	MMM	Textual	Textual+Lookback
NLH	0.764	0.723	0.894	0.917
PCH	0.775	0.416	0.941	0.945

IV. Conclusion

In this study, a multilevel accompaniment generation method for piano based on temporal data modeling is proposed, which applies the constant Q-transform and the short-time Fourier transform to the time-frequency analysis of piano audio, and designs a multilevel accompaniment generation model based on the structure of the codec. Through experimental validation on three datasets, LPD, Freemidi and GPMD, the model in this paper shows significant advantages over MuseGAN and MMM. In terms of the quality of intra-track music fragment generation, the pitch quantity metric of this paper's model on the GPMD dataset differs from the real data by only -0.445, which is better than MMM's +0.718; in terms of the inter-track harmonization metric, the inter-track distance between the piano and the guitar of this paper's model on the LPD dataset is only 0.632, which is much lower than MuseGAN's 1.174 and MMM's 1.387. Especially on the pitch and note length distribution, the note length histogram (NLH) coverage of this paper's model reaches 0.917, and the scale histogram (PCH) coverage reaches 0.945, which are higher than the comparison models. The results show that the Lookback mechanism encoding can effectively improve the quality of accompaniment generation and increase the note length histogram coverage by 0.023. The model proposed in this paper not only generates single-track high-quality accompaniment music, but also achieves good results in maintaining the harmony among multiple tracks, which provides a new technological path for generating multi-level accompaniment effects in piano art instruction, and has important theoretical value and practical significance.

References

MacRitchie, J. (2015). The art and science behind piano touch: a review connecting multi-disciplinary literature. Musicae Scientiae, 19(2), 171-190.

Dongmei, S., & Binqi, Y. (2021). The artistry in piano accompaniment teaching. Curriculum and Teaching Methodology, 4(5), 93-96.

Fang, S. (2019). The Connotation of Piano Improvised Accompaniment and the Ability Accomplishment of Accompanist. In Proceedings of the 1st Asia International Symposium on Arts, Literature, Language and Culture (AISALLC 2019) (pp. 64-69).

Tabuena, A. C. (2020). Functional approach as a beginning method for teaching and learning piano accompaniment in music. International Journal of Trend in Scientific Research and Development, 5(1), 51-54.

Vilar, J. M. P., & Grau, L. V. (2020). Vocal piano accompaniment: a constant research towards emancipation. English Language Literature & Culture, 5(1), 13-24.

Li, J. (2020). Analysis of piano curriculum education and cultivation of creative thinking ability. Region-Educational Research and Reviews, 2(1), 6-8.

Ertem, A. (2023). Piano Accompaniment Competences of Music Teacher Candidates. Cumhuriyet Uluslararası Eğitim Dergisi, 12(4), 1038-1047.

Zhifei, S. Z. S., & Pattananon, N. (2022). Development and Benefits of Teaching Piano Accompaniment Courses in Higher Music Education, China. Journal of Modern Learning Development, 7(7), 437-444.

Wei, L. (2020, November). Cultivation of Music Learning Ability in the Teaching Reform of Collective Piano Lesson. In 2020 International Conference on Social Sciences and Big Data Application (ICSSBDA 2020) (pp. 151-155). Atlantis Press.

Han, X., & Xu, M. (2024). Cultivating Diverse Harmonic Thinking in Improvisational Accompaniment: Strategies and Case Studies Exploration. Lecture Notes in Education Psychology and Public Media, 51, 197-204.

Pengtao, X., & Somtrakool, K. (2024). Piano accompaniment course in music university in China. Journal of Roi Kaensarn Academi, 9(1), 643-655.

Dong, X. (2024). The Development of Applied Piano Teaching Method to Cultivate Undergraduate Students' Ability in Music Learning Skills. Sciences of Conservation and Archaeology, 36(3), 291-302.

Xin, W., & Pattananon, N. (2022). The Important of Piano Accompaniment Skills for Teaching Music for Early Childhood in Taishan, China. Journal of Modern Learning Development, 7(9), 376-386.

Luo, J. (2024). The Application of'Improvisational Accompaniment'in Piano Textbooks for Music Majors in Chinese Higher Education. International Educational Research, 7(3), p40-p40.

Zhihang Meng, Meng Zhihang & Chen Wencheng. (2020). Automatic music transcription based on convolutional neural network, constant Q transform and MFCC. Journal of Physics: Conference Series,1651(1),012192-.

Assem Abdelhakim. (2025). Radiation anomaly detection with source identification capability using short-time fourier transform. Journal of Instrumentation,20(04),P04024-P04024.

Liu Weiming. (2022). Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition. The Journal of Supercomputing,79(6),6560-6582.