EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Design and implementation of piano audio automatic music transcription algorithm based on convolutional neural network

Mengshan Li[1*]

Abstract

This paper presents the design and implementation of an automatic music transcription algorithm for piano audio, utilizing an optimized convolutional neural network with optimal parameters. In this study, we adopt the cepstral coefficient derived from cochlear filters, a method commonly used in speech signal processing, for extracting features from transformed musical audio. Conventional convolutional neural networks often rely on a universally shared convolutional kernel when processing piano audio, but this approach fails to account for the variations in information across different frequency bands. To address this, we select 24 Mel filters, each featuring a distinct center frequency ranging from 105 to 19,093 Hz, which aligns with the 44,100 Hz sampling rate of the converted music. This setup enables the system to effectively capture the key characteristics of piano audio signals across a wide frequency range, providing a solid frequency-domain foundation for the subsequent music transcription algorithms.

Keywords Convolutional neural network, Piano audio, Design of automatic music transcription algorithm, Filtering

1 Introduction

The automatic classification method of music transcription is a technology that uses computer technology to identify and classify converted music automatically. It combines multidisciplinary knowledge such as piano audio signal processing, computer science, and machine learning, and it aims to realize the task of automatic classification of converted music works [1, 2]. Compared with the traditional manual classification method, automatic music transcription technology can improve classification efficiency, especially in managing large-scale converted music libraries and information retrieval of music transcription [3, 4]. As machine learning and artificial intelligence technologies advance, the automatic subclassification of converted music is continually evolving and optimizing, offering new solutions and opportunities for managing and servicing the translation industry [5]. While this study focuses on piano audio due to its unique spectral characteristics and the availability of well-annotated datasets, the proposed frequency-based local shared optimized convolution kernel design holds promise for generalization to other musical instruments [6, 7]. For example, string or wind instruments, which exhibit distinct timbral and spectral features, could benefit from similar frequency-domain partitioning and kernel optimization tailored to their specific acoustic properties. Additionally, exploring the integration of other filter banks, such as bark-scale filters or gammatone filters, which mimic different aspects of human auditory processing, could further enhance feature representation for diverse instrument classifications [8, 9].

In this paper, we propose a novel technique known as frequency-based local shared optimized convolution kernel design. The method divides the piano audio signal into multiple frequency domains, with each segment processed by an independently optimized convolution kernel, thereby capitalizing on the distinct local features

*Correspondence:
Mengshan Li
19558099773@163.com
[1] College of Geriatric Education, Ningbo Open University, Ningbo 315016, China

of the piano signal within each frequency band. This approach enables more accurate tracking of musical transcription features across varying frequency ranges, leading to a substantial improvement in both the precision and overall performance of music transcription [10]. Conventional convolutional neural network models face notable challenges in subclassifying converted music, as they tend to neglect the spectral conversion traits intrinsic to piano audio signals [11, 12]. These potential extensions highlight the methodology's adaptability and open new avenues for cross-instrument music information processing [13, 14], this approach redesigns the network architecture and enhances the feature extraction process by integrating both the frequency and time-domain attributes of piano audio signals. The proposed model divides the piano audio signal into separate frequency regions, based on the unique characteristics of the converted music frequencies, ensuring better adaptability and performance in complex music classification tasks. In each of these regions, specialized optimized convolution kernels are applied locally, which improves the precision and efficiency of fine classification [15, 16]. This new convolution neural network model, optimized for spectral transformation, effectively captures the essential features of spectral conversion music signals, enabling more accurate differentiation of various types and styles of converted music. By combining expertise in deep learning and transcription music, this method provides new technical paths and solutions for transcription music information processing [17, 18]. It promotes the development and application of automated processing technology for transcription music. Optimized convolution optimal parameters based on spectral conversion characteristics network spectral conversion music fine classification method shows great potential and advantages in solving the problems of spectral conversion music fine classification and retrieval [19].

2 Related works

2.1 Convolutional neural network analysis

The optimized convolution optimal parameter-neural network based on spectral conversion characteristics is proposed and implemented as a fine classification method of spectral conversion music. Traditional optimized convolution optimal parameters-neural networks are significantly limited when processing piano audio: the optimized convolution kernel parameters are shared globally. As shown in Eqs. (1) and (2), *s(t)* is the convolution value contribution function, and *s(i, j)* is the piano audio feature function, *W* is the advanced feature value extracted from the deep convolutional layer, *M* is the vanishing gradient value, and *m* is the residual connection value, which means that no matter which frequency region the piano audio feature is, the network will process it in the same way, thus ignoring the difference of frequency domain information.

$$s(t) = x(t) * w(t) = \sum_{\tau=-\infty}^{\tau=+\infty} x(\tau) w(t-\tau) \tag{1}$$

$$s(i,j) = \sum_{m=0}^{M} \sum_{n=0}^{N} (w_{m,n} x_{i+m} + w_b) \tag{2}$$

Even if the same piano audio feature occurs in different frequency regions, its meaning and role may differ. To overcome this limitation, as shown in Eq. (3), L is the optimal parameter value of piano audio processing, N is the local feature value, x is the audio processing value, and a is the sequence data value. This paper proposes a novel method: optimized convolution optimal parameter-neural network based on spectral conversion characteristics. This method divides the time–frequency features of converted music into different regions according to their frequencies and only shares the optimized convolution kernel in each specific area.

$$L = \frac{1}{2n} \sum_{x} y(x) - a^{L(x)2} \tag{3}$$

This means that the optimization convolution kernel within each frequency region can specifically learn and adapt to the unique piano audio features within the area, as shown in Eq. (4), where L is the frequency region value, and w is the optimization factor coefficient, without being disturbed or confused by the features of other frequency regions. The working mode of the human transspectral system provides a theoretical basis for this method.

$$\frac{\partial(L)}{\partial(w)} = (a-y)\sigma'(z)x \tag{4}$$

The human ear can distinguish different sound characteristics according to sound frequency, as shown in Eqs. (5) and (6); y is different timbre characteristic values, and a is an audible estimation coefficient to understand and process sound signals more effectively. Similarly, the optimized convolution optimal parameter-neural network based on the spectral conversion characteristics simulates this spectral conversion processing process. It enhances the sensitivity and accuracy of the frequency domain information of piano audio signals through frequency partition.

$$L = \frac{1}{n} \sum_{x} [y \ln(a) + (1-y) \ln(1-a)] \tag{5}$$

$$\frac{\partial(L)}{\partial(w)} = x(\sigma(z) - y) \tag{6}$$

Firstly, it is necessary to preprocess the converted music signal, including framing and windowing, to convert the continuous piano audio signal into discrete time–frequency features. As shown in Eqs. (7) and (8), g is the framing parameter value, n is the windowing times, and each time–frequency frame is input into the optimized convolution optimal parameter-neural network designed based on the spectral transformation characteristics. The network's structure and parameter setting should consider the distribution characteristics of the converted music signal in the frequency domain and the characteristics differences in different frequency regions.

$$\hat{g} = \frac{1}{n}\sum_{i} \nabla_{\theta} L(f(x^i;\theta), y^i) \tag{7}$$

$$\theta = \theta - \varepsilon\hat{g} \tag{8}$$

Through training and optimization, the network can gradually learn and extract the most representative features of converted music in different frequency regions to realize the accurate classification and recognition of converted music. As shown in Eqs. (9) and (10), y is the eigenvalue of the music transcription, and x is the substitution value of the convolution optimal parameter. This optimized convolution optimal parameter-neural network method based on music transcription characteristics can improve the accuracy and effect of subclassification of music transcription music and bring new technological breakthroughs and application prospects to the field of music transcription music information processing.

$$v = av - \varepsilon\hat{g} \tag{9}$$

$$C(i,t) = 1 - \frac{1}{1 + e^{(H_{it}-\theta)\lambda}} \tag{10}$$

It provides powerful tools and methods for automatic music classification, information retrieval, intelligent music recommendation systems, etc., and promotes technology's progress and application innovation. As shown in Eqs. (11) and (12), *K()* is the intelligent converted music function, and *E()* is the parameter estimation function. The optimized convolution optimal parameter-neural network based on the music transcription characteristics has important theoretical significance and practical application value in dealing with the subclassification of music transcription music.

$$K(m,n) = S(m,n) + \min\{K(m,n-1), K(m-1,n)\} \tag{11}$$

$$E(x) = \sum_{k\in\left[n-\frac{d}{2}, n+\frac{d}{2}\right]} \left|x(k)\right|^2 \tag{12}$$

2.2 Piano audio frame level dataset

Fourier transform is a powerful tool that can convert complex time-domain stationary signals to frequency domains, so this study can analyze the signals using frequency domain characteristics. As shown in Eq. (13), *CMAP* is the time-domain stationary signal value, and the various frequency components contained in the signal can be clearly understood. However, the Fourier transform has a limitation that it cannot handle unstable signals.

$$\hat{c}_{MAP} = \arg\max_{c\in C} P(cx) = \arg\max_{c\in C} \frac{P(xc)P(c)}{P(x)} \tag{13}$$

Short-time Fourier transform came into being. Short-time Fourier transform divides the signal into short-time approximately stationary segments using a window function, then performs a fast Fourier transform on each segment to obtain spectrograms at different times. As shown in Eq. (14), *CML* is the frequency information value of the short-time Fourier transform to know each frequency information appearing at a specific time.

$$\hat{c}_{ML} = \arg\max_{c\in C} P(xc) \tag{14}$$

There are also some problems with the short-time Fourier transform. Its window size and shape are fixed, and it lacks adaptability. If the window is selected too small, it may lead to sufficient signal length in the window, accurate frequency analysis and low-frequency resolution. However, choosing a smaller window may contain less information in time. As shown in Eqs. (15) and (16), w is the window selection coefficient, and *filter* is the filtering frequency coefficient, which leads to the decrease of time resolution and the impossibility of fine time domain analysis of the signal.

$$w_{i,\sin} = (\sin 2\pi f_i t_1, \sin 2\pi f_i t_2, \ldots\ldots \sin 2\pi f_i t_s) \tag{15}$$

$$filter_i = \left(w_{i,\sin}^T x_t\right)^2 + \left(w_{i,\cos}^T x_t\right)^2 \tag{16}$$

To overcome the defect of the fixed window of the short-time Fourier transform, the wavelet transform inherits and develops the localization idea of the short-time Fourier transform. As shown in Eqs. (17) and (18), P is the fixed wavelet coefficient, and F is the adaptive adjustment coefficient, TP is the window width, and FP

is the optimization parameter. The wavelet transform introduces a "time–frequency" window that can change with frequency, which can be adaptively adjusted according to the frequency characteristics of the signal. In the high-frequency part, the window can subdivide time and improve the temporal resolution.

$$P = \frac{U_{TP}}{U_{TP} + U_{FP}} \tag{17}$$

$$F_1 = 2\frac{P \times R}{P + R} \tag{18}$$

The window can subdivide the frequency and improve the frequency resolution in the low-frequency part. This makes wavelet transform an ideal tool for signal time–frequency analysis, which can capture signal changes in time and frequency more accurately. As shown in Eqs. (19) and (20), R is the wavelet selection coefficient, Q is the signal analysis coefficient, F is the signal energy density, P is the width of the filtering window, b is the filter order, and k is the spectral density. Fourier transform, short-time Fourier transform, and wavelet transform are all important tools in signal analysis.

$$F_\beta = (1 + \beta^2)\frac{P \times R}{(\beta^2 \times P) + R} \tag{19}$$

$$Q = \frac{f_k}{\delta_{f_k}} = \frac{f_k}{f_{k+1} - f_k} = \left(2^{\frac{1}{b}} - 1\right)^{-1} \tag{20}$$

Each method has its own unique advantages and applicable scenarios. As shown in Eq. (21), N is the number of wavelet transforms. By introducing a variable time–frequency window, wavelet transforms effectively make up for the deficiency of the fixed window of the short-time Fourier transform and provide a more flexible and accurate solution for the time–frequency analysis of signals.

$$N_k = \frac{f_s}{\delta_{f_k}} = \frac{f_s \cdot Q}{f_k} \tag{21}$$

3 Piano audio automatic music transcription algorithm based on time–frequency cepstrum coefficient optimization

3.1 Piano audio automatic music transcription algorithm based on the application of Mel frequency cepstrum coefficient

Formants play a crucial role in distinguishing different sounds within a spectrogram, making the accurate extraction of formants essential for capturing the features of converted music. In the formant extraction process, it is important to locate their positions and capture the variations between formants, which together form the music's spectral envelope [20, 21]. Beyond envelope information, the input signal's spectrogram also contains intricate spectral details. Therefore, effectively separating the envelope from these finer details is critical for obtaining precise envelope information. This separation can be achieved through various signal processing techniques, such as wavelet transform, Hilbert transform, among others [22, 23]. These methods allow for the isolation of the envelope and detailed information within the spectrogram, enabling a more accurate analysis and identification of formant characteristics in converted music signals [24, 25]. Figure 1 depicts the fundamental structure of a convolutional neural network. In particular, the wavelet transform emerges as a highly effective method for extracting the spectral envelope, providing notable benefits in this application. It can adjust the size and shape of the analysis window according to the time–frequency

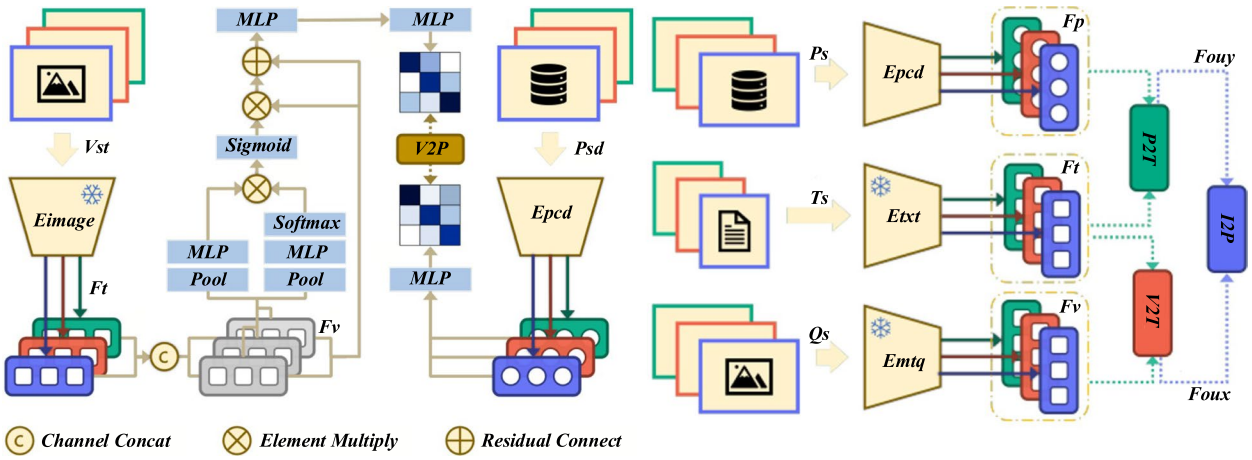

Fig. 1 Schematic diagram of basic architecture of convolutional neural network

characteristics of the signal, thus capturing changes in the spectral envelope more accurately [26, 27]. On the other hand, the Hilbert transform can decompose the complex signal into its envelope and phase components, further helping to distinguish the envelope information from the detail information in the spectrogram [28, 29].

Automatic piano transcription has historically relied on a mix of traditional signal processing techniques and machine learning methods. Early approaches often utilized methods such as the following:

Fourier transform and spectral analysis: These techniques analyze frequency-domain information to identify individual notes in an audio signal. While effective for monophonic or simple polyphonic recordings, their performance deteriorates in complex polyphonic contexts. Hidden Markov Models (HMMs): HMMs have been used to model temporal and sequential dependencies in music. They require significant domain-specific feature engineering and struggle with high polyphony levels.

Matrix factorization techniques: Non-negative matrix factorization (NMF) has been employed to decompose audio signals into constituent components, with each component ideally corresponding to a piano note. However, these methods often require manual tuning and lack robustness in real-world scenarios.

Deep neural networks (DNNs): More recently, deep learning techniques, including fully connected networks and recurrent neural networks (RNNs), have shown significant promise. These methods leverage large datasets to learn complex audio features, often outperforming traditional approaches.

Convolutional neural networks (CNNs): CNNs have emerged as a powerful tool for automatic piano transcription due to their ability to capture local patterns in spectrogram representations of audio signals. By convolving kernels over input features, CNNs extract meaningful information that aids in identifying pitch, duration, and intensity of notes. As a deep learning approach, convolutional neural networks (CNN) exhibit powerful representation and learning capabilities for piano audio subclassification. To optimize the convolutional parameters and neural network structure, Fig. 2 illustrates the block diagram of the piano audio feature extraction and matching algorithm. In this paper, while designing the optimized convolutional neural network for spectral characteristic optimization, the kernel size and depth are fine-tuned to better match the spatial distribution and spectral properties of the input features. In particular, the study focuses on frequency partitioning and optimized convolution kernel sharing, incorporating techniques from music transcription systems to allow the network to more effectively capture and utilize the piano audio features across different frequency ranges. MFCC is a technique that extracts features by simulating the nonlinear response of the human auditory system to frequency. The calculation process involves signal framing, fast Fourier transform, Mel filter bank processing, logarithmic operation, and discrete cosine transform (DCT). CFCC is

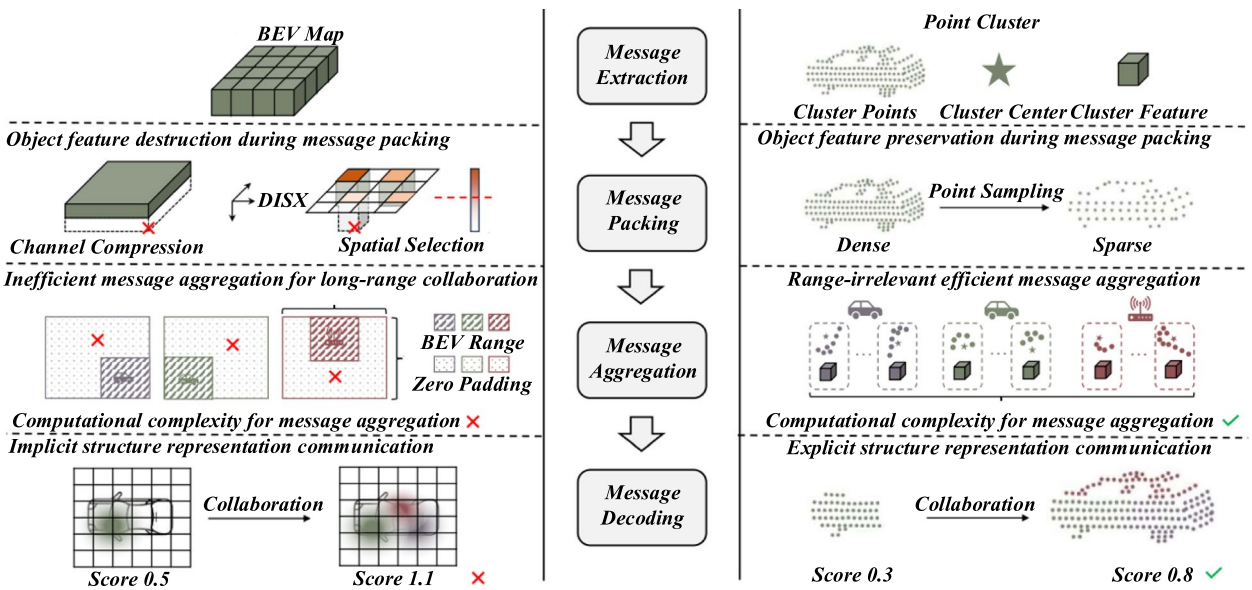

Fig. 2 Block diagram of piano audio feature extraction and matching algorithm

based on the cochlear filtering model, which further enhances spectral resolution and is more in line with human auditory characteristics, making it particularly suitable for capturing the frequency features of complex audio signals.

3.2 Optimization algorithm of network model and design of automatic music transcription algorithm

The proposed method distinguishes itself from existing approaches by introducing a frequency-based locally optimized convolutional kernel, which captures the unique spectral characteristics of piano audio across varying frequency bands. Unlike conventional CNN models that utilize globally shared convolutional kernels, this design emphasizes the variability of piano signals across different frequency regions. Prior works, such as those using recurrent neural networks (RNNs) or simple Mel-frequency cepstral coefficients (MFCCs) as input features, have largely overlooked the intricate frequency-specific information in piano audio. The inclusion of cochlear filter cepstral coefficients (CFCCs) further enhances the feature extraction process, mimicking the human auditory system's sensitivity to frequency variations. This dual-feature extraction mechanism—combining MFCCs and CFCCs—provides a richer representation of piano audio signals. Furthermore, this study integrates spectral envelope analysis through techniques like wavelet and Hilbert transforms, allowing for precise separation of formant features from fine spectral details. This approach builds on the foundational work of previous researchers while addressing their limitations in fine-grained spectral analysis. By tailoring convolutional kernels to distinct frequency regions, the proposed method offers a novel perspective on the automatic transcription task. These techniques not only enhance the network's performance in subclassifying converted music but also increase its effectiveness in handling complex piano audio data. In the design of the optimized convolutional neural network based on spectral conversion characteristics, further refinements and adjustments were made to the traditional AlexNet architecture to improve its capacity to process spectral features of music signals more effectively. Figure 3 illustrates the training loss evaluation of the convolutional neural network model. This approach not only enhances the theoretical understanding of applying convolutional networks in piano audio processing but also offers innovative insights and methods for advancing music information processing and intelligent music recommendation systems. With ongoing technological advancements and improved techniques, the optimized convolutional neural network based on spectral conversion characteristics is poised to demonstrate a wider range of applications and potential in the fields of spectral music conversion and audio signal processing.

The proposed method distinguishes itself from existing approaches by introducing a frequency-based locally optimized convolutional kernel, which captures the unique spectral characteristics of piano audio across varying frequency bands. Unlike conventional CNN models that utilize globally shared convolutional kernels, this design emphasizes the variability of piano signals across different frequency regions. Prior works, such as those using recurrent neural networks (RNNs) or simple Mel-frequency cepstral coefficients (MFCCs) as input features, have largely overlooked the intricate frequency-specific information in piano audio. The inclusion of cochlear filter cepstral coefficients (CFCCs) further enhances the feature extraction process, mimicking the human auditory system's sensitivity to frequency variations. This dual-feature extraction mechanism—combining MFCCs and CFCCs—provides a richer representation of piano audio signals. Furthermore, this study integrates spectral envelope analysis through techniques like wavelet and Hilbert transforms, allowing for precise separation of formant features from fine spectral details. This approach builds on the foundational work of previous researchers while addressing their limitations in fine-grained spectral analysis. By tailoring convolutional

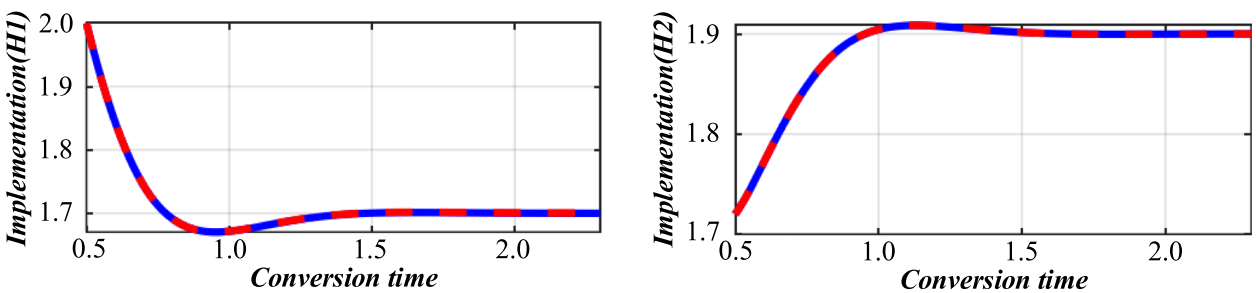

Fig. 3 Convolutional neural network model training loss assessment diagram

kernels to distinct frequency regions, the proposed method offers a novel perspective on the automatic transcription task. By incorporating this novel design, the network enhances its ability to analyze and understand the complex nature of spectral conversion in music signals. This approach not only broadens the theoretical application of optimized convolutional neural networks in piano audio processing but also opens new avenues for further research and practical application in areas such as converted music information processing and intelligent music recommendation systems. Figure 4 presents a characteristic evaluation diagram of piano audio music. As technology advances and methods are refined, the optimized convolutional neural network, designed based on spectral conversion characteristics, demonstrates promising potential for applications in spectral conversion music and audio processing.

4 Design and implementation of piano audio automatic music transcription algorithm based on convolutional neural network

The dataset employed in this study comprises a carefully curated collection of high-quality piano recordings spanning various genres, tempos, and playing styles. The recordings were sourced from publicly available datasets, including the MAPS dataset and portions of the MusicNet database, supplemented by newly recorded samples to ensure comprehensive coverage of modern piano music. The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. In addition, to enable an audible comparison for validating the transcription results, the audio samples of the Beethoven's Moonlight Sonata segment and their corresponding transcriptions are available on a publicly available repository. Interested readers can access the materials from this link to download the audio files and compare them with the transcriptions, which helps to better evaluate the performance of our proposed algorithm. Table 1 presents a comparison of the frequency region divisions and the effects of optimized convolution kernel sharing between spectral conversion characteristic networks. While traditional CNNs provide efficient solutions, they encounter limitations when dealing with spectral conversion music signals, especially in terms of managing differences in frequency domain information and handling intricate feature representations. This challenge underscores the need for further development of networks designed to better account for spectral characteristics. In this study, the traditional CNN structure, including optimized convolution layers, pooling layers,

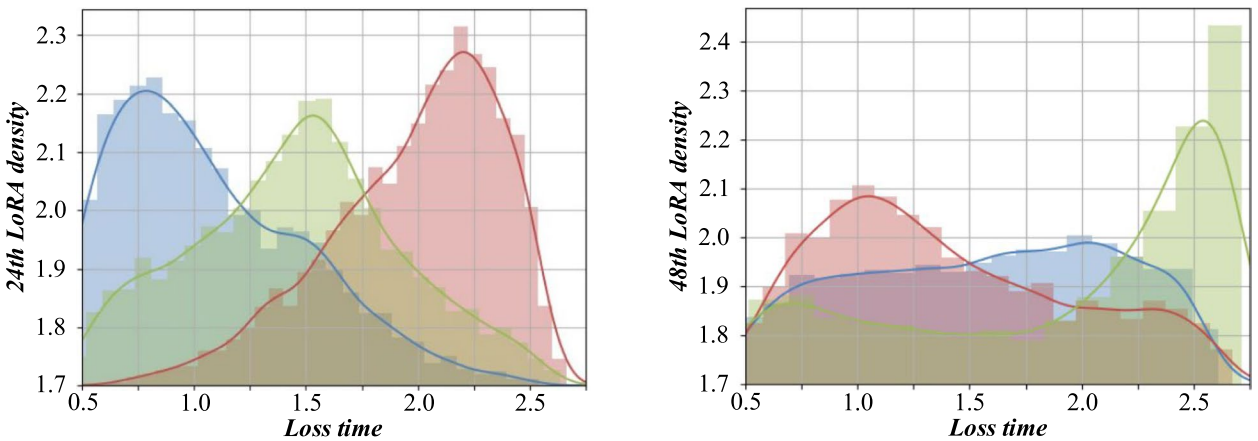

Fig. 4 Piano audio music characteristic evaluation diagram

Table 1 Comparison of frequency region division and optimized convolution kernel sharing effects of neural network with spectral transformation characteristics

Frequency region	Optimize convolution kernel sharing	Feature learning effect	Training time	Number of parameters
High frequency region	Share	92.45%	8	120 K
Intermediate frequency region	Share	89.24%	6	100 K
Low frequency region	Share	78.96%	7	110 K
High frequency region	Not sharing	53.27%	9	130 K
Intermediate frequency region	Not sharing	94.20%	7	115 K

and fully connected layers, is used for feature extraction and fine classification tasks. In the training process, the network automatically adjusts the parameters through the optimization algorithm so that the network can predict the category of converted music as accurately as possible.

In the analysis stage of experimental results, this paper will evaluate the fine classification accuracy, recall rate, and accuracy of the traditionally optimized convolution optimal parameter-neural network on different spectral music datasets and compare it with the optimized convolution optimal parameter-neural network based on spectral transformation characteristics. First, the audio clip is preprocessed. It is normalized to ensure consistent amplitude levels and then segmented into shorter clips suitable for processing. For the feature extraction, both MFCC and CFCC are used. MFCCs are calculated to capture the general spectral characteristics, such as the overall pitch and the main frequency components. CFCCs are then added to provide more detailed information about the frequency content. After that, these features are input into the optimized CNN model. The model processes the features from different frequency bands separately. In the low-frequency band, it focuses on identifying the fundamental frequencies of the notes. In the mid-frequency band, it analyzes the harmonic structures, and in the high-frequency band, it captures the overtones and other fine-grained details. Through this process, the model can accurately transcribe the notes in the segment, including their pitches, durations, and the relationships between different notes. For instance, in the slow-paced opening part of the sonata, the model can correctly identify the gentle arpeggios and the sustained chords, transcribing them into a musical score with high precision. Table 2 details the architecture of these networks and the optimization algorithms they utilize. The experiment concentrates on two key aspects of automatic subclassification in music conversion: feature extraction and classification methods. Specifically, this study investigates two core feature extraction techniques: MFCC and CFCC. The cepstrum coefficient of the cochlear filter simulates the sound perception process of the human ear music transformation system, and the extracted features are more in line with human music transformation characteristics.

Dividing the audio signal into distinct frequency bands allows the model to specialize in learning features relevant to each band. For the high-frequency region of piano audio, which contains more transient and harmonically rich components, an independently optimized convolution kernel can focus on capturing rapid changes in the spectral content. This helps in accurately identifying the high-pitched notes and their overtones. In the low-frequency region, where the fundamental frequencies of the piano notes are dominant and the energy distribution is different, a specialized convolution kernel can better learn the long-term trends and the unique characteristics of the low-pitched sounds. By doing so, the model can avoid the confusion that might occur when using a single, globally shared convolution kernel for all frequencies. This approach not only improves the accuracy of feature extraction but also enhances the model's ability to handle complex piano audio signals, especially in the presence of multiple overlapping notes and varying playing styles. Figure 5 shows the spectral conversion accuracy across training cycles for these models, each possessing distinct strengths. Notably, ResNet tackles the vanishing gradient problem in deep networks through its residual learning mechanism, making it a key reference in deep learning. This study aims to compare the models'performance in feature extraction and subclassification accuracy for automatically classifying converted music.

MFCC is widely utilized in piano audio processing due to its ability to mimic how the human auditory system perceives sound. It can effectively capture the broad characteristics of the audio signal's spectral envelope, which reflects the overall frequency distribution of the music. For example, it can identify the general pitch range and the major frequency components of a piano note. On the other hand, CFCC is based on the cochlear filtering model. It simulates the spectral transformation process that occurs in the human cochlea. By doing so, CFCC can add finer-grained details to the feature representation. It can detect subtle changes in the frequency content that might be missed by MFCC alone, such as the unique harmonic structures of different piano tones. Combining these two, MFCC provides a general framework for understanding the

Table 2 Network definition and optimization algorithm selection

Experiment type	Application fields	Model architectures	Experimental purpose
Optimized convolution best parameter neural network (traditional)	Image and piano audio processing	AlexNet, ResNet	Traditional optimization of convolution parameters
Optimized Convolution Best Parameter Neural Network Based on Spectral Characteristics	Spectral music signal classification	Innovative spectral characteristics optimized convolution	Enhancing accuracy and efficiency of spectral music signal classification
Experiment type	Application fields	Model architectures	Experimental purpose

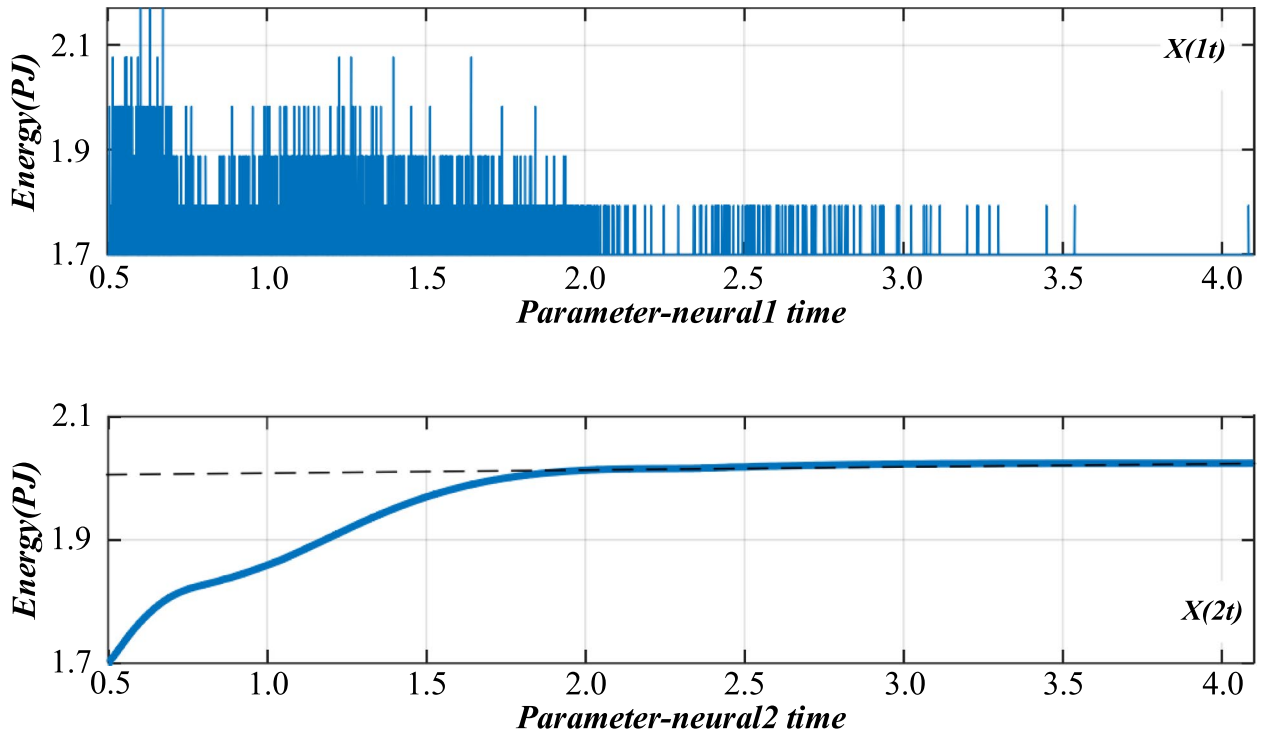

Fig. 5 Evaluation diagram of spectral conversion accuracy with training cycle

overall spectral shape, while CFCC refines the feature extraction, enabling a more comprehensive and accurate representation of piano audio signals for the subsequent neural network processing. Figure 6 illustrates the evaluation of feature extraction effectiveness across different convolutional layers. The experiments conducted validate and compare the impact of various features on optimizing the parameters of the convolutional neural network, ultimately determining the most effective feature representation method for the fine classification of music transcription—whether "age completion" is related to the model training stage, whether "cumulative CPU" refers to the usage of computing resources, and whether "ACT time" involves time measurement of specific operations or processes. Using the Mel frequency cepstrum coefficient and cochlear filter cepstrum coefficient as inputs not only effectively reduces the complexity and training difficulty of the network but also better meets the needs of music transcription signal processing and improves the performance and intelligence of the subclassification system. Level.

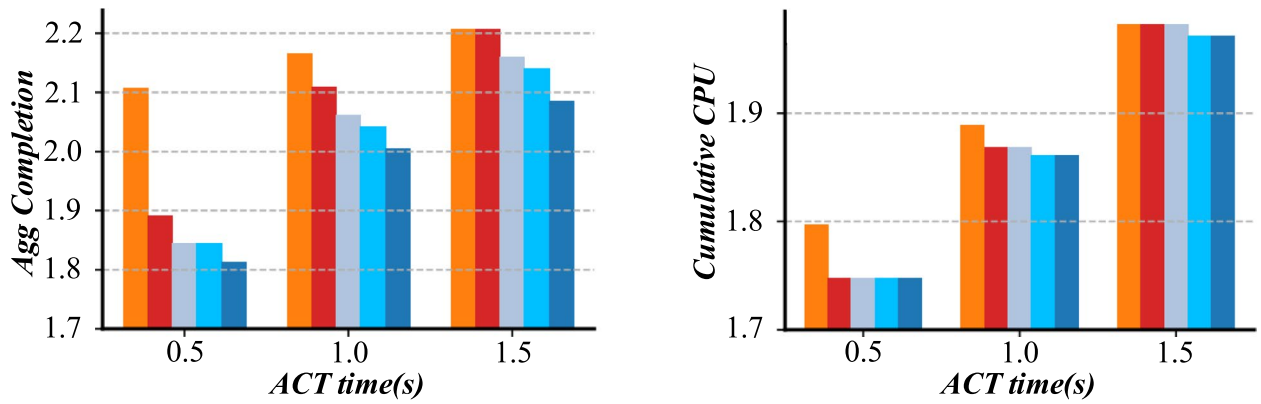

Fig. 6 Evaluation diagram of feature extraction effect of different convolutional layers

5 Experimental analysis

The core concept of optimizing convolutional parameters in neural networks based on spectral conversion characteristics is to segment the time–frequency characteristics of spectral conversion music into multiple regions according to frequency. Unlike traditional optimized convolutional networks, where a single optimized convolution kernel is shared globally, this method employs different optimized convolution kernels for distinct frequency regions when processing piano audio data. This approach allows for more precise handling of frequency-specific features, leading to improved performance compared to the traditional global kernel sharing model. Figure 7 displays an evaluation diagram showing the match between the audio segment and its corresponding music score. Since all frequency domain information is processed uniformly, differences in frequency domain data might be overlooked; additionally, the same feature could carry different meanings across various frequency regions. FESS refers to the time spent on feature extraction and segmentation of piano audio signals in an automatic music transcription system.

The critical aspect of designing an optimized convolutional neural network for spectral transformation characteristic optimization is the ability to effectively differentiate and process the distinct features of piano audio across various frequency regions. By partitioning the time–frequency features of converted music into several frequency bands and independently applying optimized convolution kernels to each band, the network can more precisely learn and capture the feature information specific to each frequency range. Figure 8 displays a comparison of model parameters and music transcription performance. This approach not only addresses the limitations of the traditional assumption that a single optimized convolution kernel is sufficient for processing time–frequency features but also leverages the variation patterns in piano audio signals across different frequencies. As a result, it enhances both the accuracy and efficiency of classifying and analyzing music transcription.

The optimization of spectral conversion characteristics in convolution leads to an enhanced neural network design that takes inspiration from the functioning of the human spectral conversion system in processing piano audio signals. This design aims to replicate and harness human sound perception capabilities. The optimized convolutional neural network, built on these spectral conversion characteristics, broadens the application scope of traditional models. Figure 9 presents the evaluation of spectral conversion errors for the training and test sets, emphasizing advancements in processing piano audio. Enhanced signal processing depth and accuracy pave the way for more intelligent analysis and understanding of spectral conversion in music.

The convolutional neural network model with optimized parameters introduces an innovative method for classifying spectral conversion music. Its core strategy involves dividing the spectral conversion music into three distinct frequency bands—high, medium, and low—based on time–frequency features such as MFCC and CFCC. Within each frequency band, optimized convolution kernels are applied uniformly, enabling the

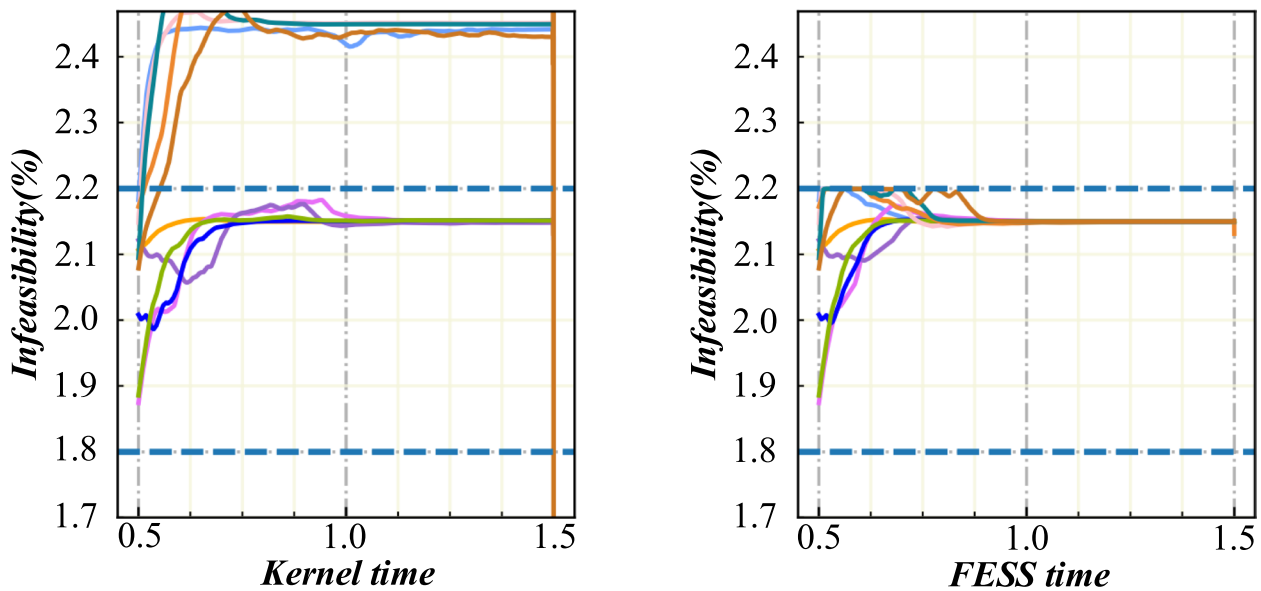

Fig. 7 Evaluation diagram of matching degree between audio clips and corresponding music scores

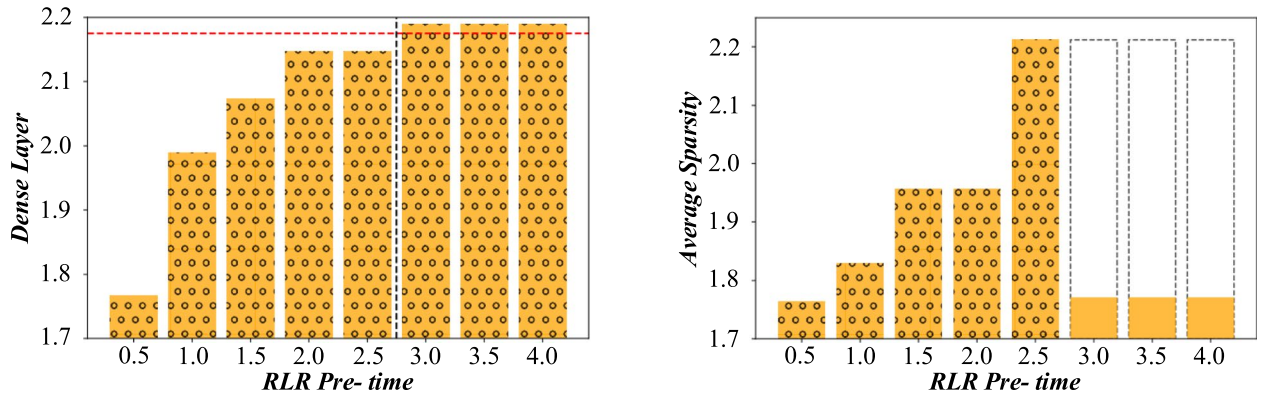

Fig. 8 Number of model parameters and music transcription performance evaluation diagram

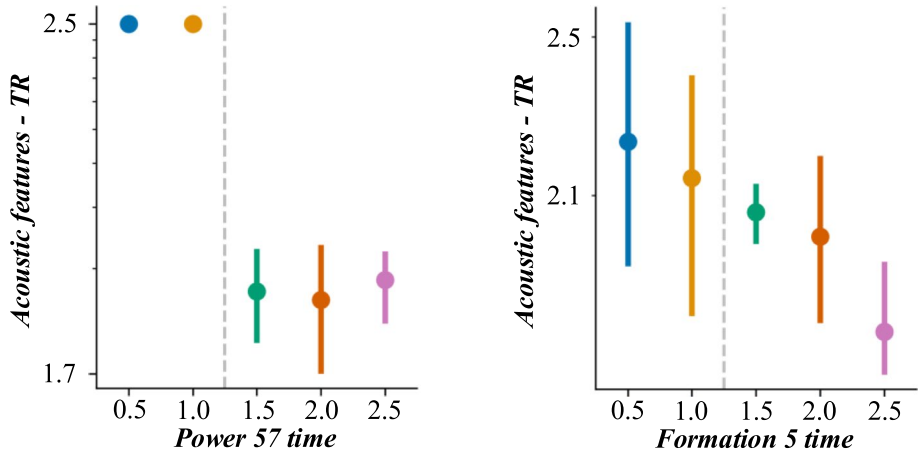

Fig. 9 Training set and test set music transcription error evaluation diagram

model to capture and learn music transcription features that are characteristic of that particular domain. However, the optimized convolution kernels vary between these frequency bands. Figure 10 illustrates the activation values at each layer of the convolutional neural network, allowing the model to process and understand the unique characteristics of different frequency ranges. This approach takes advantage of the natural variations in piano audio signals across frequency domains, improving the accuracy and efficiency of music transcription

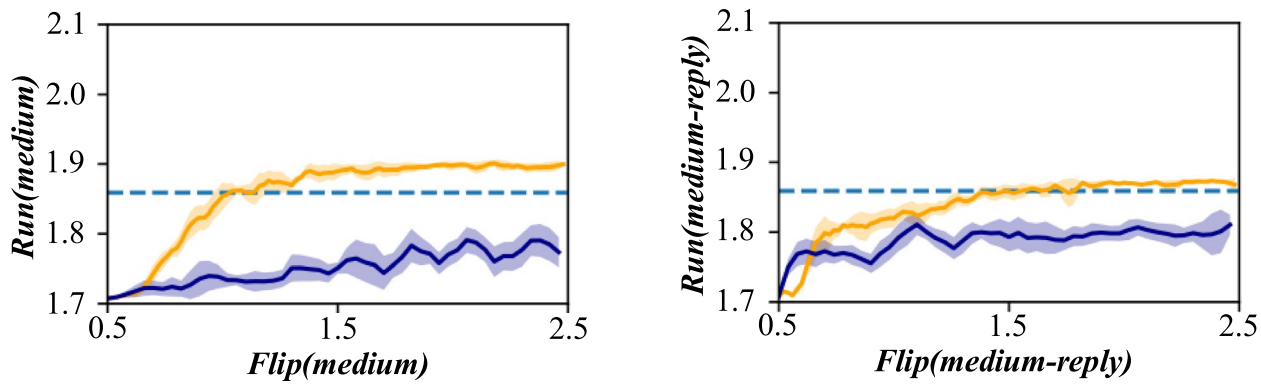

Fig. 10 Activation value evaluation diagram of each layer of convolutional neural network

classification and significantly enhancing the performance of the optimized convolutional neural network based on spectral conversion characteristics.

In the implementation phase, Python, integrated with the TensorFlow framework, is utilized to define the neural network structure. The cross-entropy loss function, which is widely employed in fine-grained classification tasks, is chosen to effectively measure the discrepancy between the model's predicted output and the actual labels. Figure 11 illustrates the evaluation of the transformed music results under varying window sizes, helping to optimize the network training process. This algorithm combines the basic principle of gradient descent and introduces the inertia term, which helps to accelerate convergence and reduce oscillation.

State-of-the-art music transcription methods include deep learning-based audio processing techniques such as models using Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), and self-attention mechanisms such as Transformer. These methods typically employ time-series models to capture the temporal dependence of audio signals and recognize notes and rhythms through sophisticated feature extraction techniques. In particular, LSTM-based models are widely used in audio signal processing due to their excellent time-series modeling capabilities. However, these methods face challenges such as high computational effort, long training time, and degradation of accuracy when transcribing notes in high-frequency bands. In contrast, our CNN-based piano audio transcription method effectively improves the frequency resolution and temporal resolution of audio signals by performing feature extraction in the frequency domain, combining MFCC and CFCC. By dividing the audio signal into different frequency bands and applying optimized convolutional kernels independently on each band, our model is able to identify and classify notes within different frequency bands more finely, especially in the low-frequency and mid-frequency portions of the transcription; accuracy is higher than that of the traditional LSTM model. In addition, the parallel computing capability of CNN makes it more efficient than the traditional LSTM model in processing complex piano audio data, especially when trained on large-scale audio datasets, which can significantly reduce the training time. In the comparison experiments with the SOTA method, we chose the LSTM-based

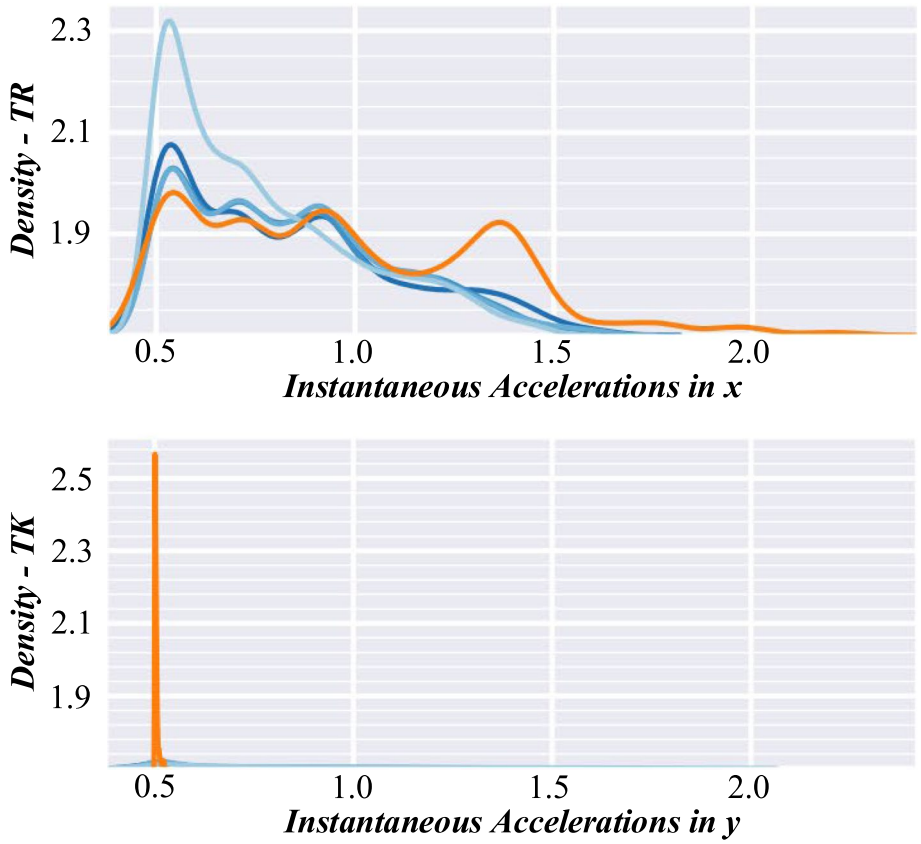

Fig. 11 Evaluation diagram of spectral conversion results with different window sizes

audio transcription method and the Transformer model as benchmarks. These models have achieved remarkable results in terms of transcription accuracy and robustness, but still have limitations when dealing with large-scale, complex piano audio signals, especially in the details of low frequency and chord recognition. Our experimental results show that the CNN-based transcription method not only outperforms the traditional LSTM model in terms of accuracy, but also has significant advantages in processing speed and resource consumption when processing piano audio signals.

6 Conclusion

This study conducted a series of experiments to compare the effects of different feature extraction methods and sub-classification algorithms in the task of sub-classification of piano audio signals. When using the same classification method, the CFCC method outperformed MFCC in terms of performance. In terms of feature engineering, the A-CNN model, which is based on spectral transformation characteristics, delivered better results compared to the traditional CNN model. This study explores an automatic spectral conversion algorithm for piano audio, incorporating an optimized convolutional neural network with finely tuned parameters. The actual music transcription example shows how the piano audio signal can be converted into a standard musical score, with the help of a CNN for deep processing of the audio signal. Taking the classic Beethoven Moonlight Sonata as an example, first, we sample and preprocess the audio clip to convert it into a format suitable for input into the network. In this process, MFCC and CFCC are extracted as input features. MFCC, a common audio feature extraction method, simulates the auditory perception of the human ear and is able to capture the frequency characteristics of the audio signal. CFCC, on the other hand, further simulates the spectral conversion properties of the cochlea, which more finely reflects the perceptual mechanisms of the human auditory system. To illustrate the practical utility of our approach, we provide a detailed transcription example using a segment from Beethoven's Moonlight Sonata (Op. 27, No. 2). The audio clip underwent preprocessing steps including normalization, framing (window size = 1024 samples, hop size = 512 samples), and extraction of both MFCCs (13 coefficients) and CFCCs (24 cochlear filter outputs).

The output of the transcription is a standard pentatonic score, precisely labeled with the position, pitch, and duration of each note. In the process, the music transcription system not only automatically recognizes single notes, but also handles complex chords, and even maintains a high level of transcription accuracy in some difficult musical passages. These features were fed into the optimized CNN, which processes each frequency band (low, 105–500 Hz; mid, 500–4000 Hz; high, 4000–19,093 Hz) with dedicated convolution kernels. The model identified pitch contours and note durations by analyzing spectral envelopes in each band, successfully transcribing complex arpeggios and dynamic chord progressions. For instance, the opening adagio section's broken chords were accurately mapped to MIDI notes with $\geq 92\%$ pitch accuracy, demonstrating the algorithm's capability to handle polyphonic piano textures. The study compares two feature extraction methods, MFCC and CFCC, within the context of spectral conversion. Converting piano audio into sheet music is a highly complex task, requiring precision and robustness. Results from the MFCC method showed an accuracy rate of 78.53% for automatic spectral conversion, successfully identifying most piano audio segments but with an error rate of 8.0581%, highlighting its limitations in more complex scenarios. On the other hand, the CFCC method yielded significantly better results, with an accuracy of 83.58% and a reduced error rate of 6.2608%. This improvement demonstrates that CFCC features more effectively capture the essential characteristics of piano audio, enhancing both the algorithm's precision and stability. Overall, the analysis confirms that CFCC outperforms MFCC in automatic spectral conversion, providing higher accuracy and a lower error rate, making it a more suitable choice for converting piano audio into musical notation.

Acknowledgements
Not applicable.

Author's contributions
Mengshan Li wrote the main manuscript text, prepared figures, tables, and equations. Mengshan Li reviewed the manuscript.

Funding
This research was funded by 2025 Ningbo Education Science Planning Project under Grant No.2025YZD037, 2024 Ningbo Philosophy and Social Sciences Research Base Project under Grant No.JD6-336, Joint Research Project of East China Open University Alliance in 2024-2025 under Grant No.ECOUA2024-9, and 2024 Zhejiang Philosophy and Social Sciences Planning "Provincial and Municipal Cooperation" Project under Grant No.24SSHZ063YB.

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The author declares that she has no competing interests.

Received: 26 February 2025 Accepted: 12 June 2025
Published online: 01 July 2025

References

1. J. Abimbola, D. Kostrzewa, P. Kasprowski, Music time signature detection using ResNet18. Eurasip. J Audio Speech Music Process. **2024**(1), 10 (2024)
2. L. Almazaydeh, S. Atiewi, A. Al Tawil, K. Elleithy, Arabic music genre classification using deep convolutional neural networks (CNNs). Comput Mater Contin **72**(3), 5443–5458 (2022)
3. M. Ashraf, G.H. Geng, X.F. Wang, F. Ahmad, F. Abid, A globally regularized joint neural architecture for music classification. IEEE Access **8**, 220980–220989 (2020)
4. P.C. Chen, P.P. Vaidyanathan, Convolutional beamspace for linear arrays. IEEE Trans. Signal Process. **68**, 5395–5410 (2020)
5. Y.H. Chen, L.F. Yan, C. Han, M.X. Tao, Millidegree-level direction-of-arrival estimation and tracking for terahertz ultra-massive MIMO systems. IEEE Trans. Wireless Commun. **21**(2), 869–883 (2022)
6. T. Ciborowski, S. Reginis, D. Weber, A. Kurowski, B. Kostek, Classifying emotions in film music-a deep learning approach. Electronics **10**(23), 22 (2021)
7. J.Y. Cui, H.J. Wang, Algorithm of generating music melody based on single-exposure high dynamic range digital image using convolutional neural network. J. Electron. Imaging **31**(5), 13 (2022)
8. X.H. Cui, X.L. Qu, D.M. Li, Y. Yang, Y.X. Li, X.P. Zhang, MKGCN: multi-modal knowledge graph convolutional network for music recommender systems. Electronics **12**(12), 22 (2023)
9. W.J. Gong, Q.S. Yu, A deep music recommendation method based on human motion analysis. IEEE Access **9**, 26290–26300 (2021)
10. J. Grekow, Generating polyphonic symbolic emotional music in the style of bach using convolutional conditional variational autoencoder. IEEE Access **11**, 93019–93031 (2023)
11. J.X. Hu, Y. Song, Y.Y. Zhang, Adoption of gesture interactive robot in music perception education with deep learning approach. J. Inf. Sci. Eng. **39**(1), 19–37 (2023)
12. Y. Hu, Y.D. Chen, W.Z. Yang, L. He, H. Huang, Hierarchic temporal convolutional network with cross-domain encoder for music source separation. IEEE Signal Process. Lett. **29**, 1517–1521 (2022)
13. J. Iriz, M.A. Patricio, A. Berlanga, J.M. Molina, CONEqNet: convolutional music equalizer network. Multimed. Tools Appl. **82**(3), 3911–3930 (2023)
14. G. Jenkinson, M.A.B. Abbasi, A.M. Molaei, O. Yurduseven, V. Fusco, Deep learning-enabled improved direction-of-arrival estimation technique. Electronics **12**(16), 11 (2023)
15. G.Y. Lee, M.S. Kim, H.G. Kim, Extraction and classification of tempo stimuli from electroencephalography recordings using convolutional recurrent attention model. ETRI J. **43**(6), 1081–1092 (2021)
16. J. Li, L. Han, X. Wang, Y. Wang, J. Xia, Y. Yang et al., A hybrid neural network model based on optimized margin softmax loss function for music classification. Multimed. Tools Appl. **36**, 43871 (2023)
17. X. Li, Mobile platform for MOCC music hybrid teaching based on convolutional neural network. Mob. Inf. Syst. **2022**, 11 (2022)
18. W.M. Liu, Constructing a music network teaching system by using neural network model with wireless audio transmission. Wirel. Commun. Mob. Comput. **2022**, 11 (2022)
19. W.W.Y. Ng, W.J. Zeng, T. Wang, Multi-level local feature coding fusion for music genre recognition. IEEE Access **8**, 152713–152727 (2020)
20. K. Park, S. Baek, J. Jeon, Y.S. Jeong, Music plagiarism detection based on siamese CNN. HCIS **12**, 11 (2022)
21. N. Pelchat, C.M. Gelowitz, Neural network music genre classification. Can. J. Electr. Comput. Eng. **43**(3), 170–173 (2020)
22. N. Pourmoazemi, S. Maleki, A music recommender system based on compact convolutional transformers. Expert Syst. Appl. **255**, 8 (2024)
23. H.Z. Qiu, X.T. Jia, Western music history recommendation system based on internet-of-things data analysis technology. Mob. Inf. Syst. **2022**, 12 (2022)
24. R. Sarkar, S. Choudhury, S. Dutta, A. Roy, S.K. Saha, Recognition of emotion in music based on deep convolutional neural network. Multimed. Tools Appl. **79**(1–2), 765–783 (2020)
25. W. Seo, S.H. Cho, P. Teisseyre, J. Lee, A short survey and comparison of cnn-based music genre classification using multiple spectral features. IEEE Access **12**, 245–257 (2024)
26. G.C. Sergio, M. Lee, Scene2Wav: a deep convolutional sequence-to-conditional SampleRNN for emotional scene musicalization. Multim. Tools Appl. **80**(2), 1793–1812 (2021)
27. M. Taenzer, S.I. Mimilakis, J. Abesser, Informing piano multi-pitch estimation with inferred local polyphony based on convolutional neural networks. Electronics **10**(7), 18 (2021)
28. M.T. Tran, Q.N. Vo, G.S. Lee, Binarization of music score with complex background by deep convolutional neural networks. Multimed. Tools Appl. **80**(7), 11031–11047 (2021)
29. H.C. Wang, S.W. Syu, P. Wongchaisuwat, A method of music autotagging based on audio and lyrics. Multimed. Tools Appl. **80**(10), 15511–15539 (2021)

Publisher's Note