Dean&Francis
ISSN 2959-6157

Artificial Intelligence Approaches in Music Audio Analysis

Weihan Wang

School of Mathematical Sciences, Beijing Normal University, Beijing, China
*Corresponding author: 202311130015@mail.bnu.edu.cn

Abstract:

As people's pursuit of art gradually increases, the forms and types of music have become increasingly diverse. In the process of exploring different types of music, the demand for music audio analysis has gradually increased. With the continuous development of artificial intelligence technology, machine learning, deep learning and other methods have gradually entered the public eye with their efficient data processing capabilities. Therefore, artificial intelligence methods in music audio analysis have gradually become the focus of research. At present, the processing mode based on traditional features such as Mel frequency cepstral coefficients (MFCC) combined with machine learning methods has been widely adopted, but there are still certain limitations in feature expression ability and classification accuracy. In recent years, end-to-end deep learning methods have shown stronger adaptability and accuracy by automatically extracting features for classification and recognition, promoting the advancement of pure music audio analysis technology. This article aims to provide theoretical support and practical guidance for researchers in related fields by organizing the application of artificial intelligence methods in pure music audio analysis, comparing and analyzing the advantages and disadvantages of various methods, and promoting the sustainable development and technological innovation of this field.

Keywords: Mel Frequency Cepstral Coefficients (MFCC); Machine Learning Methods; Audio Analysis; Feature Extraction.

1. Introduction

With the rapid development of machine learning technology, music audio analysis, as an important branch of audio processing, has gradually become a research focus. The processing and classification of music audio is not only of great significance in applications such as music information retrieval and music recommendation systems, but also promotes continuous innovation in related algorithms and models.

Music audio analysis is dedicated to the complete process from signal acquisition, feature extraction, to high-level semantic understanding. In recent years, with the widespread application of artificial intelligence methods, significant progress has been made in this field. Deep learning models represented by convolutional neural networks and recurrent neural networks not only improve the efficiency of audio feature extraction, but also promote the improvement of task performance such as music emotion recognition, sound quality evaluation, and personalized recommendation.

However, there are still many challenges in the analysis process of pure music signals. Due to the diversity of audio sources and different acquisition devices, the generalization ability of models in actual scenes is often affected, and domain adaptation has become an urgent problem. Meanwhile, differences in audio compression and encoding formats may also interfere with the stability of acoustic features. In addition, the high cost and uneven quality of music data annotation constrain the performance of supervised learning models, and the reliability and applicability of techniques such as synthetic data augmentation still need further verification.

This paper systematically reviews the methods and advancements of artificial intelligence in music audio analysis, with a focus on core technologies such as feature learning, model construction, and cross-domain adaptation. It summarizes existing issues and prospects future research directions, providing valuable references for related research and applications.

2. Audio Features

Audio features are digital representations extracted from the original audio signal, typically described in terms of time domain, frequency domain and perceptual dimensions: time domain features such as zero-crossing rate and energy reflect the macroscopic properties of the signal's amplitude over time; frequency domain features such as spectral centroid and Mel Frequency Cepstral Coefficients (MFCC) reveal the energy distribution of the sound in the frequency domain and the texture of the timbre; perceptual features such as loudness and brightness map the physical signal to the subjective auditory perception of the human ear. These diverse features enable the quantification and analysis of abstract musical elements such as rhythm, harmony and timbre, providing an information basis for subsequent intelligent models.

This study focuses on introducing the important audio feature in the field of music audio analysis - Mel frequency cepstral coefficients. Mel Frequency Cepstral Coefficients is a widely used feature extraction method in audio signal processing, particularly in the fields of speech recognition, sentiment analysis, and music information retrieval. Its core is to simulate the perceptual characteristics of the human auditory system towards sound frequency, converting the linear frequency scale to the Mel frequency scale, so as to make the extracted features more in line with the auditory perception rules of the human ear.

The specific processing method is shown in Figure 1 below.

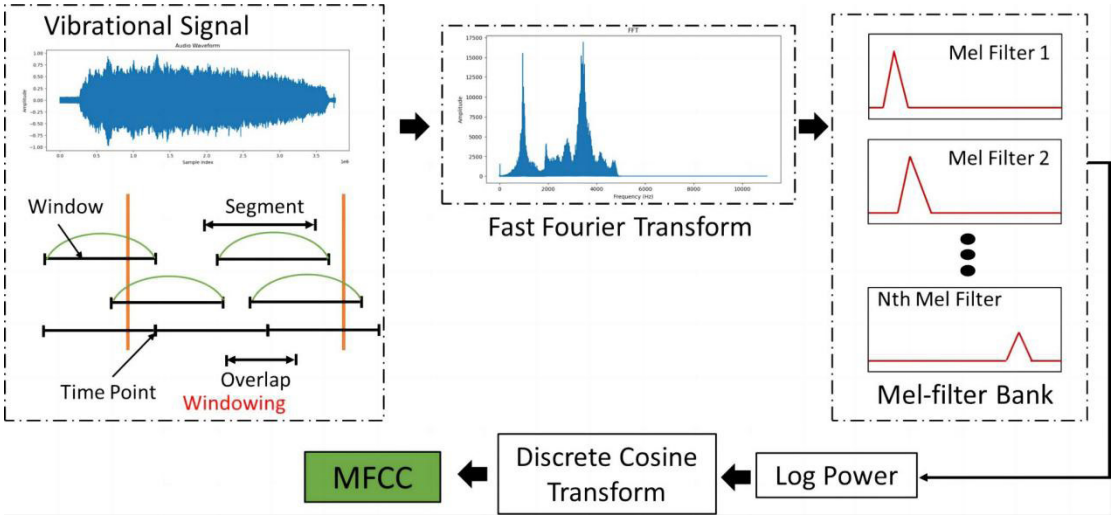

Fig. 1 Mel frequency cepstral coefficient calculation flowchart [1]

By dividing the original audio signal into frames, applying a windowing function to each frame, calculating its short-time Fourier transform to obtain the spectral energy distribution, and then mapping the spectrum to a filter bank on the Mel frequency scale for weighted summation, the Mel frequency energy is obtained. The Mel scale formula is:

$$M(f) = 1125 \cdot \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

ISSN 2959-6157

Finally, take the logarithm of these energies and perform a discrete cosine transform to obtain the final cepstral coefficients. It is precisely because MFCC can effectively capture the short-term frequency characteristics and timbre information of sound that it has strong discriminative ability.

3. Audio Based Instrument Recognition Method

3.1 Machine Learning Based Methods

In the field of music audio analysis, Mel frequency cepstral coefficients are a widely used feature extraction method that can effectively capture the spectral characteristics of audio signals, simulate the sensitivity of the human auditory system to different frequencies, and extract key features of music. Support Vector Machine (SVM), as a common machine learning classification algorithm, is often used in audio signal classification tasks due to its good generalization ability in high-dimensional space and adaptability to small sample data. Therefore, the combination of MFCC and SVM has become a classic combination in audio analysis, especially demonstrating high effectiveness in the recognition and classification of music audio.

Support Vector Machine is a very popular and powerful supervised learning algorithm, mainly used for classification problems, and can also be used for regression (known as Support Vector Regression). Its core is to label all positive and negative samples in one space, and find a separating hyperplane in this space that maximizes the distance between it and the positive and negative samples. Taking two-dimensional space as an example, the idea is shown in the following Figure 2。

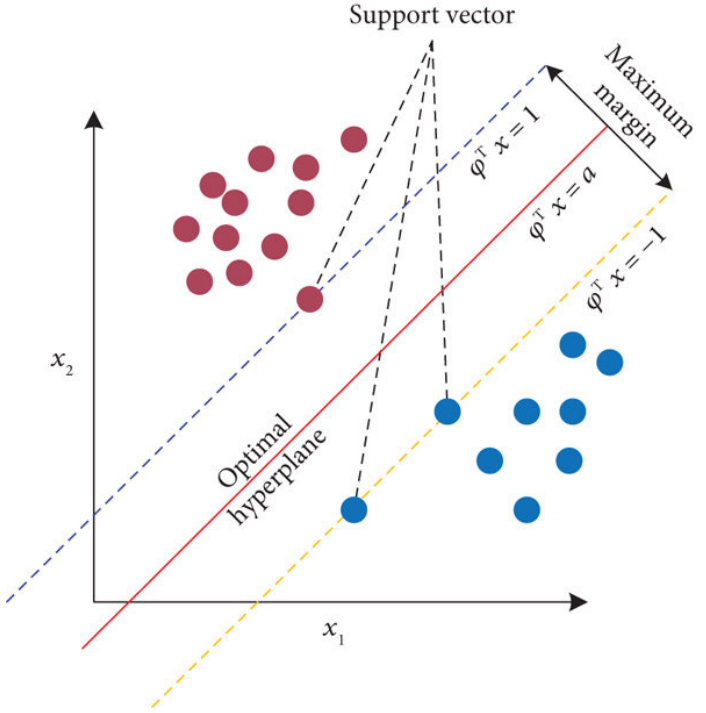

Fig. 2 Schematic diagram of the SVM model [2]

The core idea is to find the optimal separating hyperplane (as illustrated by the red line in Fig. 2) that maximizes the margin between classes, a concept that generalizes to high-dimensional spaces.

In the music audio analysis research discussed in this article, the Mel frequency cepstral coefficient audio feature extraction method is used to transform the original audio data into a 13 dimensional vector, and combine it with the support vector machine method to search for a classification hyperplane in the 13 dimensional space for classification, thus obtaining the classification standard data. Half of the overall dataset is used for training, while the other half is used as the test set for classification testing.

Through literature review, it was found that the method of combining Mel frequency cepstral coefficients with support vector machines has a classification and recognition accuracy of over 90%, which is significantly better than classification methods such as naive Bayes classification and linear regression [3].

Deep learning with use of MFCC features and conven-

tional machine learning in the music field has also performed excellently. Reportedly, in the case of electric guitar performance skill recognition MFCC was combined with other spectral features and subsequently classified with SVM that obtained an accuracy of 99.12 percent [4]. Moreover, in the area of speakers recognition, the hybrid machine learning system of MFCC and SVM also attained more than 97-percent recognition, which indicates the usefulness of the method in the voice identity recognition [5] [6].

The literature reveals that the use of the MFCC and SVM is also able to give drastic classification with the use of only these two. As an illustration, the combined features of MFCC together with SVM classifier produced a classification accuracy of 91.95 when used in the process of detecting the abnormalities in the heart sounds indicating the efficiency of MFCC features in the sound signals and the efficient classification capacity of SVM [7]. Equally, when it comes to respiratory phase detection, the optimally set MFCC parameters with the use of SVM classifier also increased the classification rate up to 87.16% [8]. Besides, MFCC/SVM combination has demonstrated excellent recognition capability in insect sound classification and speech disease identification [7] [9].

Mel Frequency Cepstral Coefficients with Support Vector Machines integration has shown to have great success in classification accuracy and in real-world applications in a variety of fields including biological sound recognition, disease diagnostics, and analyzing music performances. The results confirm the effectiveness of MFCC as an audio feature and the strong classification ability of SVM, and it provides a theoretical basis and practical model of music audio analysis. Future application in the combination of this approach with multi-feature fusion and ensemble learning has potentials of wider use and better results in the area of music audio analysis.

3.2 Methods Based on Deep Learning

Deep learning specifically refers to machine learning based on deep neural network models and methods. It is developed by combining statistical machine learning, artificial neural networks, and other algorithm models with contemporary big data and computing power. Deep learning has the ability to automatically extract features, and the automatically extracted features have stronger and more robust representation capabilities compared to manually designed features.

In recent years, deep learning methods have been widely applied and rapidly developed in the field of audio analysis. Their multi-level computational models can automatically extract high-order audio features from audio signals, which to some extent overcomes the limitations of traditional methods in feature extraction and expression ability. Traditional audio analysis methods typically rely on manually designed audio features, such as Mel frequency cepstral coefficients, spectral centroids, zero crossing rates, etc., and use shallow machine learning models (such as support vector machines, random forests) for classification or regression. This type of method performs well on specific tasks, but often exhibits poor robustness and generalization ability when facing complex polyphonic audio, non-stationary signals, and environmental noise. In contrast, deep learning learns audio feature representations in an end-to-end manner, avoiding the subjective limitations of manually designed features and effectively capturing the spatiotemporal dynamic information of audio signals, thereby improving the accuracy of audio recognition and classification.

We take the spectrogram as the initial feature of music audio as an example, and use the deep neural network for music audio processing as an example to discuss[10]. The core idea of using deep neural network to process music audio task is to convert audio signal into image representation, and then use the deep learning model which is excellent in the image field to learn and classify features. This process is mainly divided into three key stages: Audio preprocessing and spectrum generation, depth feature extraction, and classification and optimization. First, the audio needs to be converted into a spectrum diagram, and the spectrum is calculated by short-time Fourier transform and other methods, so as to generate a two-dimensional spectrum diagram that can contain as much original audio information as possible. This transforms the audio classification problem into an image classification problem.

Secondly, the pre trained deep convolutional neural network is used for feature extraction. The research of the Chinese side in this article has migrated vggnet, RESNET, densenet and other pre training models on Imagenet. These models can automatically learn multi-level features from the spectrum: the shallow convolution may capture the basic time-frequency texture and edge (such as the beginning and end of notes, harmonic structure), while the deep network can abstract high-level semantic features (such as the mode, rhythm or emotional tone of a specific instrument). For example, resnet50 effectively alleviates the gradient disappearance problem of deep network through its residual block and jump connection structure, so it can learn more complex and more discriminative feature representation, which is the key to its general superiority over other models in experiments.

Finally, the network end is classified through the full connection layer and softmax function. An important innovation of this study is to propose a balanced credible loss

ISSN 2959-6157

function to solve the problem that softmax cross entropy loss easily leads to over fitting and output polarization (prediction probability is too confident). This function takes the fusion of the traditional one hot tag and a uniform distribution vector as the learning goal, and adjusts it by introducing a trust factor, so that the model prediction will not be too close to the extreme value, so as to improve the generalization ability and classification stability. The whole process constitutes an end-to-end learning system: input the original audio clip, output its audio category, and all features in the middle are automatically optimized and learned by the network.

In addition, end-to-end deep learning has also demonstrated its powerful performance in the field of music audio analysis.In summary, end-to-end learning, as an important technical route for music audio analysis, directly maps raw audio to task targets through deep neural networks, demonstrating the advantages of automation, high efficiency, strong generalization ability, and wide applicability. In the future, combining multimodal information fusion, attention mechanisms, and pre trained models, end-to-end learning will play a more critical role in the field of pure music analysis, to solve the performance bottleneck of traditional methods in complex scenarios, and achieve more intelligent and accurate music understanding and application [11] [12].

4. Dataset and Mature Audio Recognition Cases

4.1 Dataset Related to Music Audio Processing

5.1.1 Neural synthesizer

Neural Synthesizer (NSynth) is an audio dataset containing 305979 notes, each with a unique pitch and timbre. For 1006 instruments from a commercial sound source library, a four second monaural 16kHz audio clip (referred to as a note) is generated by covering all pitches (21-108) of a standard MIDI piano and five different intensity values (25, 50, 75, 100, 127). Each note lasts for three seconds before sounding and naturally decays at the last second.

Due to some instruments being unable to play all 88 pitches, each instrument can present an average of 65.4 pitches. In addition, there are occasional repeated tones between different intensity levels in commercial audio packages, resulting in an average of 4.75 independent intensity variations for each pitch.

5.1.2 Instrument recognition in musical audio signals

Instrument Recognition in Musical Audio Signals (IRMAS) is a widely used benchmark dataset in the field of music information retrieval, especially in instrument recognition research. Its data mainly includes 11 instrument types, solo and ensemble audio from different eras of the last century, thus exhibiting diversity in audio quality.

5.1.3 CCMusic

CCMusic is a Chinese music database specifically designed for music information retrieval research, with the aim of providing professional and copyrighted Chinese music resources. The CCMusic database collects audio materials of hundreds of Chinese ethnic musical instruments, including popular music and ethnic music, recorded separately and with clear copyright. A significant feature of this database is the separation of singing and accompaniment, which is of great significance for many research directions in music information retrieval. The database is recorded by professional teachers and students from the Conservatory of Music in a controlled environment, with professional limitations on recording environment, equipment, and processes, ensuring high quality and standardization of data, clear copyright, and ease of academic research use.

4.2 Mature Cases of Music Audio Recognition and Classification

Kugou Music's top ranked beat extraction scheme in MIREX 2025 is an end-to-end model composed of Audio Encoder, MLP feature mapping layer, and Transformer temporal prediction module. This model consists of three modules working together. The Audio Encoder is responsible for extracting deep features from the original audio signal, while the MLP layer maps the features to standardized Audio Tokens. Finally, with the powerful long-range context modeling capability of Transformer networks, the model can accurately predict beat sequences from Audio Tokens sequences. The excellence of this technology lies in its ability to handle complex rhythms. By conducting multi task joint training on audio datasets containing a large number of complex rhythms, the scheme performs excellently in terms of rhythm tracking accuracy, especially in identifying mixed rhythms, demonstrating significant performance advantages and generalization ability.

5. Current Limitations and Future Prospects

Data diversity and annotation quality are fundamental factors that determine the performance and generalization ability of machine learning models when analysing pure music audio. The existing paradigms of multimodal music information retrieval are using different traditional

and deep-learning models to analyze sentiment, different genres, recommender systems, and emotion recognition, but again these systems strictly rely on high-quality representative datasets. The current collections are mostly limited to a particular linguistic or cultural and the Sotho-Tswana music video dataset with both text and visual and audio modalities as well as annotations provided by native speakers provides a cultural richness. However, the low-resource languages and the specialized musical genres have a catastrophically small number of separately annotated resources [13].

Multi-class classification in pure music audio analysis presents substantial challenges in accuracy and generalization. Music signals have complex time-frequency properties of small overlappant acoustic characteristics between categories ascribing difficulties to model discrimination. In the task of recognizing individual techniques of guitar playing, such as in identifying guitar sounds, the amount of spectral features that are extracted to a nearby set of lightweight CNNs, such as MobileNetV2, InceptionV3, and ResNet50 has enhanced the recognition of nine types of guitar sounds [4]. In spite of these, the generalization of the model of real-world data is no more than 70.9 percent, and the performance of the model has shown a downgrade on the presence of more elaborate background noise and variation in performance.

Multi-class audio classification therefore faces interconnected problems such as complexity of high-dimensional features, scarcity of data, and noise in the environment, so that accuracy-generalization trade-offs are especially hard to achieve in the practice. To overcome them, researchers are trying to consider them with their diversified feature fusion, transfer learning, and reinforcement learning to be more robust and applicable, setting the key research priorities to be used in the future within this sphere.

6. Conclusion

The application of artificial intelligence methods in the field of music audio analysis has become increasingly diverse with the development of the times. From traditional machine learning methods such as combining Mel frequency cepstral coefficients with support vector machines, it has gradually evolved into an end-to-end automatic extraction of audio features for classification and recognition in deep learning. Its application scope has gradually expanded from a single music audio processing to medicine, society, education and teaching, and so on. The accuracy, robustness, and generalization ability of audio analysis have also become the main directions of development and research in related fields.

The continuous development of artificial intelligence methods in the field of pure music audio analysis not only promotes the iterative updating of technology, but also has a profound impact on music science research and related applications. In the future, emphasis should be placed on interdisciplinary integration, fully utilizing the cross disciplinary advantages of computer science, musicology, and cognitive science to promote the collaborative progress of theory and practice.

References

[1] Manjunath K, Tewary S, Khatri N, et al. In-Process Monitoring and Prediction of Machining Status in Ultraprecision Diamond Turning Using Mel-Frequency Cepstral Coefficient Approach Combined with Machine Learning. International Journal of Interactive Design and Manufacturing, 2025.

[2] Cortes Corinna, Vapnik Vladimir. Support-vector networks. Machine Learning, 1995, 20(3): 273–297.

[3] Lu Lifei. Research and Implementation of a Music Analysis and Retrieval Platform Based on Score Generation. Shanghai: Shanghai Jiao Tong University, 2019.

[4] Chellamani G K, N A, C C, et al. SpectroFusionNet: A CNN Approach Utilizing Spectrogram Fusion for Electric Guitar Play Recognition. Scientific Reports, 2025, 15: 16842.

[5] V K, S S P. Hybrid Machine Learning Classification Scheme for Speaker Identification. Journal of Forensic Sciences, 2022, 67: 1033–1048.

[6] Costantini G, Cesarini V, Brenna E. High-Level CNN and Machine Learning Methods for Speaker Recognition. Sensors, 2023, 23(7): 3461.

[7] Hosseinzadeh M, Haider A, Malik M H, Adeli M, Mzoughi O, et al. Enhanced Heart Sound Classification Using Mel-Frequency Cepstral Coefficients and Comparative Analysis of Single vs. Ensemble Classifier Strategies. PLOS ONE, 2024, 19(12): e0316645.

[8] Zhantleuova A K, Makashev Y K, Duzbayev N T. Optimizing MFCC Parameters for Breathing Phase Detection. Sensors, 2025, 25(16): 5002.

[9] Wei J-Q, Wang X-Y, Zheng X-L, Tong X. Stridulatory Organs and Sound Recognition of Three Species of Longhorn Beetles (Coleoptera: Cerambycidae). Insects, 2024, 15(11): 849.

[10] Li, J., Han, L., Li, X. et al. An evaluation of deep neural network models for music classification using spectrograms. Multimed Tools Appl 81, 4621–4647 (2022).

[11] Pandeya Y R, Bhattarai B, Lee J. Deep-Learning-Based Multimodal Emotion Classification for Music Videos. Sensors, 2021, 21(14): 4927.

[12] Jiang S, Shi N, Liu C. Analysis of Artificial Intelligence Knowledge Graphs for Online Music Learning Platform Under Deep Learning. Scientific Reports, 2025, 15: 16481.

[13] Oguike O E, Primus M. Multimodal Music Genre Classification of Sotho-Tswana Musical Videos. IEEE Access, 2025, 13: 28799–28808.