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

Application of Computer-Aided Design and Multimodal Fusion in Music Beat Detection

Jie Tian

School of Education and Music, Hainan Vocational University of Science and Technology, Hainan, Haikou, 571126, China, music.tian_bj@zit.edu.cn

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

Abstract. This article explores the application of computer-aided design (CAD) and multimodal fusion technology in music beat detection, proposing an innovative detection model. To achieve this, we utilize the GTZAN dataset for experiments, analyzing the impact of different network structures on music beat detection and classification tasks. We compare our method with several prevalent machine learning algorithms in music beat detection to comprehensively evaluate its accuracy. During the experiment, the dataset is scientifically partitioned into training, validation, and testing sets, with non-repeated random sampling technology employed to ensure data independence and non-intersection. The results indicate high accuracy in detecting various music beats, including rock, classical, and electronic dance music. The average beat detection accuracy in the test set reaches 95%, surpassing the previous method by approximately 10%. Furthermore, the average beat detection time per music sample is 20 milliseconds, demonstrating the model's high computational efficiency.

Keywords: Computer-Aided Design; Multi-Modal Fusion; Music Beat Detection
DOI: https://doi.org/10.14733/cadaps.2025.S3.26-38

1 INTRODUCTION

Beat, as one of the basic elements of music, not only supports the skeleton of music but also leads the emotional ups and downs of the audience. Traditional music beat detection mainly depends on manual labelling and simple algorithms. Although this method can realize beat recognition to a certain extent, it is often limited by complex and changeable music styles and rhythm patterns, and it isn't easy to achieve high accuracy and wide applicability [1]. Especially in the face of multiple music cultures and complex music structures, the limitations of traditional methods become more and more prominent. Therefore, exploring more intelligent and efficient beat detection means has become an important topic in the field of music information processing [2]. CAD, as an innovative method combining computer science and engineering technology, has demonstrated its powerful analysis and design capabilities in many fields. Introducing CAD into music beat detection can deeply analyze the music signal by using advanced algorithm models and also improve the efficiency of beat detection by automatic and intelligent means [3]. This change is expected to break through the limitations of traditional methods and bring new tools for music analysis. With the continuous increase in the number and accuracy of sensors, the field of human-computer interaction has ushered in new changes, and multi-mode interaction modes have emerged. These innovative models not only bring users a more natural and smooth interactive experience but also provide additional strategies for designing innovative user-friendly systems [4]. In the field of music editing, this transformation is particularly significant. Music editing software often appears complex and difficult for novice users to master, which hinders their creative process. Taking music beat detection as an example, the multi-mode interaction system of MIMOSE Music Arrangement Table Editor provides a gesture and speech-based multi-mode wrapper for music editor applications [5]. To address this issue, some researchers have proposed a wrapper-based strategy that can easily map keyboard shortcuts to multimodal actions, bringing a new interactive experience to music editing software. Traditional user computer interaction often relies on tedious keyboard and mouse operations, while multimodal interaction fully utilizes natural communication channels such as gestures and voice, making the interaction process more intuitive and efficient. The wrapper converts these multimodal actions into mouse clicks or keyboard presses, achieving seamless interaction between users and software [6].

Users no longer need to rely on traditional buttons and mixing pads to create music but can manipulate the software through gestures and music terminology keywords, just like an orchestra conductor. In terms of application, the wrapper was applied to open-source music editing software tools and a detailed evaluation was conducted. It is worth noting that this encapsulated application may not necessarily be open source, but it can capture and process events sent by users through channels different from the keyboard and mouse [7]. We evaluated the accuracy and F1 score of each interaction mode separately, and the results showed that multimodal interaction performed well in both aspects. These events are triggered by multimodal actions rather than traditional button operations, providing users with a more ecologically tuned and immersive interactive environment [8]. More importantly, the wrapper of the MIMOSE core has a wide range of applicability, making it easy to adapt to other types of applications with minimal coding work. User evaluation relies on specially customized QUIS and SUSES questionnaires, and the results show that from the perspectives of technical quality and usability, the packaging of MIMOSE core has shown encouraging performance. In addition, real users were invited to evaluate the application through the wrapper extension to verify its usability. This means that in future development, this multi-modal interaction strategy will be applied to more fields, bringing users a more convenient and efficient interaction experience [9]. In actual creation, composers often do not set specific goals when performing orchestration tasks. The OOQ framework is expected to play an important role in music beat detection, sound quality analysis and optimization, and more innovative applications, promoting the deep integration and development of music creation and technology. The uniqueness of the framework lies in its clever borrowing of the concept of digital signal processing to achieve this idea [10]. In fact, the core of a large category of orchestration exercises lies in converting music materials to enhance or weaken their specific sound quality attributes. For example, making the score sound brighter, blurry, or dense [11]. From the perspective of OOQ, score and sound are no longer seen as goals to be achieved, but as "source materials" to be processed, which coincides with the processing method of vocal cords in modern digital audio workstations [12]. Furthermore, analyzes how it collaborates with other modes (such as language, actions, etc.) in multimodal interaction to jointly construct and convey specific meanings. Music beat detection, as an important aspect of music analysis, is of great significance for a deeper understanding of how music plays a role in multimodal contexts. Through rhythm detection, we can more accurately grasp the rhythm and rhythm of music [13]. Therefore, this article incorporates music beat detection into the framework of multimodal critical discourse analysis, in order to more comprehensively reveal the meaning construction mechanism of music in multimodal contexts.

Multimodal data, encompassing audio, video, text, and other forms of information, can comprehensively reflect the diverse dimensions of music. By effectively integrating these different modes of data, we can capture beat characteristics more thoroughly, enhancing the robustness and generalization ability of beat detection. Particularly in complex and dynamic music scenes, multimodal fusion technology is anticipated to exhibit unique advantages. While CAD and multimodal fusion technology theoretically offer robust support for music beat detection, their practical effectiveness still requires rigorous experimental verification.

This article mainly uses deep learning methods to improve the accuracy of music emotion recognition and simplify the computational difficulty in the recognition process. The focus is on using multimodal information feature fusion and neural networks as improved methods and embedding them as part of the modules into the intelligent music system. This article mainly introduces the historical background and current research status of music emotion recognition at home and abroad and focuses on the mainstream methods of feature extraction and model building for discrete emotion space and continuous emotion space. And provided a detailed introduction and summary of the main research content of this article, and finally introduced the chapter arrangement of the entire text. Introduced the relevant theories of the music emotion recognition model required for this article. Starting from preprocessing, we will introduce methods for extracting emotional features from music. Finally, we will introduce commonly used emotion recognition models and analyze their advantages and disadvantages. Related research on music emotion recognition in continuous emotional space. A detailed introduction was given to the CAD model constructed in this article, the selection of the database, the fusion process of multimodal information, and finally, simulation experiments were conducted on the model constructed in this chapter. This module is an improvement based on the original WLDNN_GAN model, adding a Self Attention module to optimize the calculation process in emotion recognition, introducing relevant music databases and conducting simulation experiments. Compare the experimental results with the mainstream music emotion recognition models currently in use. Embedded the designed CAD model into the model, realizing the function of music emotion recognition, and added music generation, management, and retrieval modules to improve the entire music system.

This study is expected to provide new ideas for the field of music beat detection and promote the further application of CAD and multimodal fusion technology in music analysis. At the same time, I hope this study can bring new enlightenment and tools to music creation, teaching and other fields, and promote the inheritance of music culture.

Firstly, this study discusses the basic principles and methods of CAD in detail and analyzes its unique advantages in music beat detection. Secondly, the characteristics and fusion methods of multi-modal data are deeply studied, and how to effectively integrate different modal data into the beat detection task is explored to improve the robustness and generalization ability of detection. Then, the combination strategy of CAD and multimodal fusion technology is discussed, and an accurate and efficient beat detection model is constructed. Finally, experimental design and result analysis are carried out, representative music data sets are selected for experiments, and the performance of beat detection of different methods is compared to verify the effect of the proposed method in practical application.

2 RELATED WORKS

Faced with increasingly diverse music cultures and complex music structures, the limitations of traditional beat detection methods are becoming increasingly prominent, making it difficult to meet the requirements of high precision and efficiency. The rhythm detection status is determined by the output of language and text sentiment analysis, ensuring the comprehensiveness and accuracy of the detection results. In order to further optimize the performance of emotion recognition systems, Liao [14] proposed an innovative multi-modal beat detection model based on speech and text. On this basis, deep neural networks will also be applied to the learning and classification of fused features to achieve more accurate beat detection. At the same time, in order to capture emotional information in the text, the model also uses an efficient bidirectional long short-term memory network to extract text features. Further analysis shows that the multimodal model proposed by Liao and Gui [15] has higher recognition accuracy than the single modal model on the test dataset, and also outperforms other published multimodal models. The experimental results show that the overall recognition accuracy of text is improved by 6.70%, and the recognition accuracy of beat detection is increased by 13.85% compared with the single pattern. This achievement not only provides new ideas and methods for the field of music information processing but also lays a solid foundation for future research on beat detection. Meanwhile, the successful application of this model further demonstrates the broad application prospects of multimodal fusion in music sentiment analysis, beat detection, and other fields. Multiple pattern fusion experiments were conducted on the IEMOCAP database to verify the effectiveness of the proposed model. This significant improvement fully demonstrates the advantages of multimodal models in rhythm detection.

Maba [16] believes that combining images with music is a form of music visualization aimed at deepening understanding and perception of music information. In order to explore the emotional information in music and images fully, some scholars have innovatively incorporated the emotional classification loss function into the loss function. Firstly, the basic concepts of music visualization were outlined, and then convolutional neural networks and long short-term memory networks were combined to achieve pairing and visualization of music and images. After simulation experiments, they found that the improved deep learning music visualization algorithm showed the highest matching accuracy when the weight of the emotion classification loss function was set to 0.2. On this basis, further introduction was given to music beat detection technology. Through beat detection, we can more accurately capture the rhythm information in music. Compared with traditional keyword-based matching methods and unimproved deep learning music visualization algorithms, Pandeya and Lee's algorithm [17] can more accurately match images corresponding to music emotions. It not only helps to improve the accuracy of music and image matching but also adds richer levels and details to music visualization. This achievement not only proves the effectiveness of innovative methods in the field of music visualization but also further highlights the important role of music beat detection technology in improving the accuracy of music and image pairing.

Traditional music beat detection methods often overly rely on data labels, mainly focusing on frequency domain or physical features, while the technology proposed in this article has significantly improved performance. These features not only intuitively reveal the composition of musical instruments, but also keenly capture the changes in emotions in music, bringing new perspectives to the field of mixed instrument component analysis. Xu and Zhang [18] proposed an innovative sparse feature extraction method that combines sparse decomposition techniques and specialized dictionaries for multiple instrument components. It is intended to overcome the limitations of existing methods in identifying and analyzing mixed instrument music data components. In this digital age, the widely spread online and offline music videos provide rich materials for rhythm detection and analysis. In practical applications, users on music and video platforms often add metadata containing advanced semantics such as emotions to selected products, which provides potential application scenarios for automatic emotion analysis. To validate the validity of the dataset, Yang and Nazir [19] tested four unimodal convolutional neural networks (CNNs) and four multimodal CNNs for music and video. Firstly, fine-tune each pre-trained unimodal CNN and test its performance on invisible data. To address this challenge, we have carefully constructed a balanced music video beat dataset that covers the diversity of regions, languages, cultures, and instruments, laying a solid foundation for in-depth research. In addition, a CNN-based one-dimensional music emotion classifier was trained using raw music waveform data further to explore the complex relationship between music emotions and beats. Through these experiments, the effectiveness of the dataset in rhythm detection tasks was not only verified but also the potential of CNN in music emotion classification was demonstrated. This study not only provides new ideas and methods for the field of beat detection but also opens up new avenues for the study of music emotion analysis.

As an innovative method combining computer science and engineering technology, CAD has demonstrated its powerful analysis and design capabilities in many fields. Its core is to use the efficient computing power of computers to assist human beings in carrying out complex design tasks. The basic principle of CAD is to digitize and model the traditional design and analysis process and achieve more efficient and accurate results through computer simulation and optimization.

3 MULTI-MODAL FUSION TECHNOLOGY

Music is an important carrier for expressing emotions, so studying music emotions has become a focus of attention. With the development of research, music emotion recognition technology has been widely applied in various fields. Text mode, acoustic A-mode, and visual A-mode all use the Common Bi GRU network for sentiment regression prediction. In addition, acoustic B-mode using Wav2vec2 audio pre-training and visual B-mode using VggFace visual pre-training have been added. Each modality adopts two training and testing methods, one is to use a separate modality for training and testing. Another approach is to use fixed multimodal labels to supervise and predict multimodal emotions.

From the analysis of the results, first comparing the modalities, it can be found that the text modality is easier to accurately predict emotions, while the non-text modality is difficult to predict accurately, and the acoustic modality has a clear history. However, supervision and testing can achieve better performance, indicating that capturing emotions from modalities using a unified multimodal label to supervise the performance of each single modal emotion. To be able to sense clues for this. In summary, the current mainstream models perform poorly in acoustic unimodal tasks, revealing the challenges faced by emotion-specific acoustic feature extraction in future research. In addition, the performance difference between single modal annotation and multimodal annotation confirms that multimodal annotation may mislead the representation learning process of single modal. The structure of a multi-modal neural network for music beat detection is shown in Figure 1.

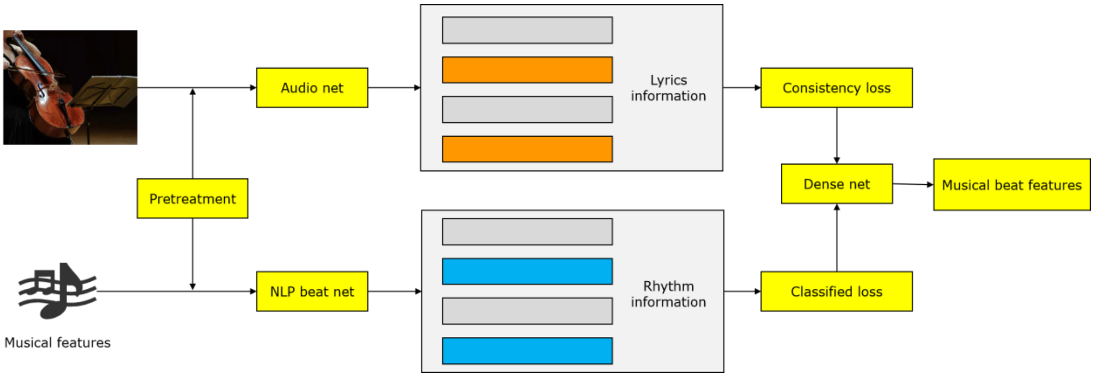

Figure 1: Multi-modal neural network structure.

4 MUSIC BEAT DETECTION BASED ON CAD AND MULTIMODAL FUSION

This article proposes a CAD music beat recognition model based on continuous emotional space and monitors music beats. Firstly, the two-dimensional emotional space Valence Acoustic and the deep learning module used to construct the model will be introduced, and the advantages and disadvantages of the model will be analyzed. Construct the WLDNN_GAN emotion recognition model, and finally fuse the extracted features into the input. In the end, a detailed introduction was given to the dataset, experimental environment, evaluation indicators, feature extraction process, and parameter design used in the experiment. And compared horizontally with mainstream models, it shows that the constructed model has better predictive performance in music emotion recognition.

The innovation of this chapter lies in the combination of MFCC and PLP acoustic features when treating music as an audio signal for data processing, which ensures the effective extraction of emotional features by extracting MFCC features. Simultaneously utilizing PLP features to enhance noise robustness, and its effectiveness has been verified through experiments. This overlapping strategy not only helps to maintain continuous rhythm information between frames but also effectively avoids beat misjudgment caused by sudden changes between frames. Especially in the case of complex and changeable music rhythms, proper sample overlap can ensure that the model is more stable when dealing with inter-frame transition, thus improving the overall beat detection effect. As shown in Figure 2, the smoothness of inter-frame transition can be clearly seen through reasonable sample overlapping design.

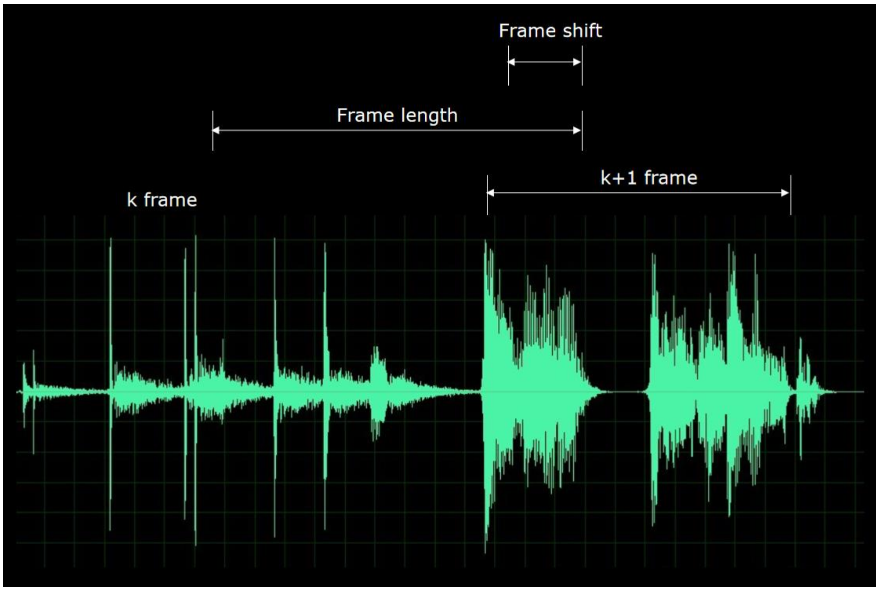

Figure 2: Overlapping framing.

After preprocessing the music samples, it becomes essential to extract a multitude of attributes pertinent to the specific identification and classification task. Subsequently, these extracted feature quantities are trained and modelled, utilizing the established extraction model to identify and classify various test music samples and obtain prediction results. Next, each frame undergoes Fourier transform, with the outcomes of each frame then stacked along an additional dimension to produce a signal resembling an image. Ultimately, the frequency axis is transformed from a linear scale to a Mel scale for dimension reduction, employing logarithmic scale values.

Regarding musical beats, the count of frames into which they can be segmented is determined as follows:

$$N_{\text{frames}} = \left[\frac{N_x - N_0}{N_f - N_0}\right] \tag{1}$$

To determine the pertinent parameters, let N_x represent the total length of the signal, N_0 denote the overlapping length between frames and N_f signify the length of one frame.

In music beat analysis, commonly employed window functions include the rectangular window, Hamming window, and Hanning window. Definitions of these window functions are provided as follows:

Rectangular window:

$$w\ n\ = 1, 0 \le n \le N-1 \tag{2}$$

Hanning window:

$$w\ \ n\ = 0.5\left(1-\cos\left(2\pi\frac{n}{N-1}\right)\right), 0\le n\le N-1 \tag{3}$$

Given N the frame length of a music beat, various window functions exert distinct impacts on the analysis of characteristic parameters. Hence, the choice of window functions should hinge on the specific characteristics of music beat parameters to effectively extract the essential features of the music.

Furthermore, due to the interrelationship between the sampling period $T_s = 1/f_s$, window length N, and frequency resolution Δf:

$$\Delta f = \frac{1}{NT_s} \tag{4}$$

It is evident that, with a constant sampling period T_s, an increase in window size N leads to a decrease in frequency resolution Δf. This implies that as frequency resolution improves, time resolution correspondingly diminishes.

$$x_i\ \ n\ = \frac{x_i\ \ n}{\max\left[x_i\ \ n\ \right]} \tag{5}$$

Here, $x_i\ \ n$ denotes the i sampling point of audio data $x\ \ n$, while $\max\left[x_i\ \ n\ \right]$ signifies the maximum value within sequence $x\ \ n$. The normalized audio data is confined to a numerical range of [-1,1].

Utilizing the spatial structure of convolution feature maps as input, each column is sequentially input into an LSTM for learning feature vectors. These feature maps are directly input into the LSTM for hash code learning. Additionally, a series of feature maps are derived from multiple convolution layers of a pre-trained CNN, considering both spatial details and semantic features. A novel loss function is designed to regulate the output of the hash layer, minimizing quantization error in the basic hash code while preserving semantic similarity and maintaining hash code balance.

While the primary task is music beat detection, preprocessing converts it into a spectrogram, framing it as an image recognition problem. Each convolution layer's feature map employs bilinear interpolation and a similarity selection strategy to form a feature map sequence, which is then input into LSTM and hash layers, ultimately recognized and classified by softmax. The proposed music beat detection framework is illustrated in Figure 3.

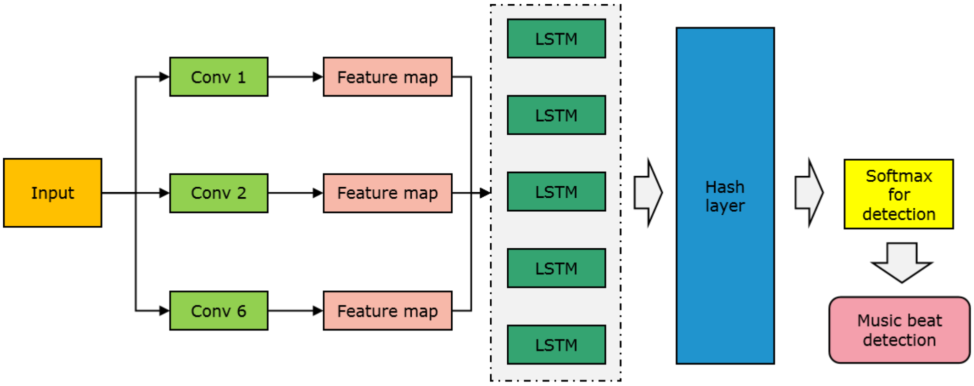

Figure 3: Music beat detection based on LSTM.

The pooling layer acts as a down-sampling layer, effectively reducing parameters and mitigating overfitting to enhance results. ReLU, a prevalent activation function in CNNs, is defined as follows:

$$f\ \ x\ \ = \max\ \ 0,x \tag{6}$$

The introduction of semi-linearity grants ReLU efficient calculation, effective gradient propagation, life probability, and a sparse activation structure, all while retaining simplicity, making it the chosen activation function for the experiment.

In the performance detection algorithm, signal stability diverges after equal iterations for both pitch and transient components, allowing for the identification of the transient part and pinpointing the signal's starting point. The redundant dictionary employs a real function set.

$$g_{\gamma}\ \ t\ \ = \frac{K_r}{\sqrt{s}} g\left(\frac{t-u}{s}\right)\cos\ \ \xi t+\varphi \tag{7}$$

When utilizing a Gaussian window g and setting K_r to $\left\|g_{\gamma}\right\|=1$, the time-frequency atoms' resolution is adjusted using a scaling factor N_{new} S, enhancing adaptability to signals. The energy distribution forms an ellipse on the time-frequency plane, with axes aligned with time and frequency. The aspect ratio of this ellipse is adjustable via scale factor s.

Discriminant training focuses on ensuring the correct label path sequence scores higher than competitors under the existing model's decoding conditions. Thus, the objective function is optimized using criteria rooted in the minimum Bayes risk framework.

$$F=\sum\nolimits_{r}\sum\nolimits_{s} p\ \ s\middle|x_r\ \ A\ \ s,s_r \tag{8}$$

$A\ \ s,s_r$ measures the accuracy of the recognized result sequence s relative to the target sequence s_r.

Superparameters, which are the initial settings for the LSTM model prior to training, differ from parameters refined through dataset learning. Their judicious selection and optimization significantly impact the training process and ultimate outcomes. To minimize the objective function, employ the generalized gradient descent method:

$$\theta_{t,i}=\theta_{t-1}-\frac{\alpha}{\sqrt{\sum_{\tau=1}^{t} g_{\tau,i}^{2}}}\, g_{t,i} \tag{9}$$

The formula includes α, which stands for the model's original learning rate, and g_{τ}, represents the decreasing gradient of i parameter $\theta_{t,i}$, with step size τ. This ratio allows for the stochastic removal of certain hidden layer neurons and their associated input-output weight parameters, ensuring neural network propagation and gradient changes remain efficient and accurate.

To harness the benefits of CAD technology and multimodal fusion, this study introduces a music beat detection model integrating both. The model initially employs CAD technology for preprocessing and extracting audio features from music signals. Additionally, it gathers modal data like video and text from these signals, extracting corresponding video and text features. During feature extraction, the model leverages signal processing in CAD and feature extraction methods in multimodal fusion to ensure the extracted features accurately reflect the music's rhythm information.

To preserve semantic tag-related information in the music's feature representation, akin to the self-attention mechanism used in related work, this article opts to concurrently calculate multiple sets of attention weights on the feature sequence, as illustrated in Figure 4.

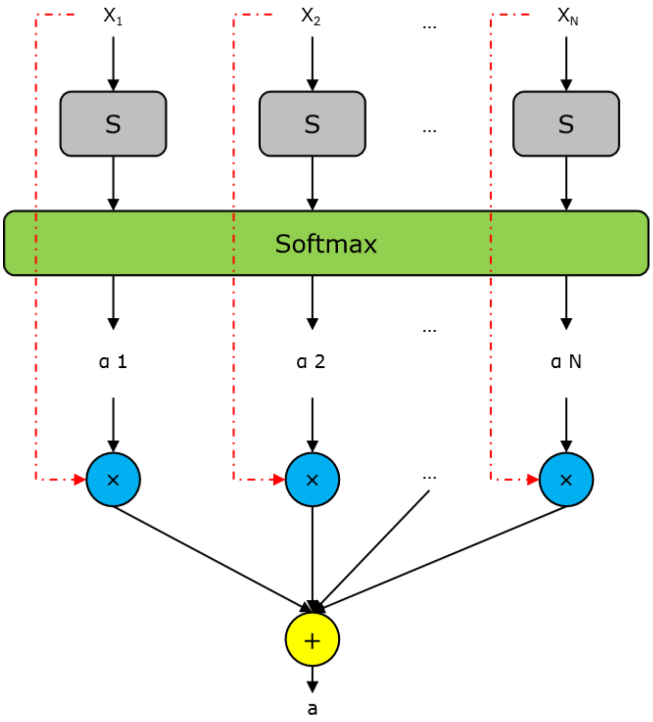

Figure 4: Schematic diagram of self-attention mechanism.

In the modal fusion stage, the model will adopt an innovative fusion strategy to effectively fuse audio features, video features and text features. This fusion strategy can fully consider the complementarity and redundancy between different modal data to extract more comprehensive and accurate beat features. Finally, in the stage of beat recognition, the model will use a deep learning algorithm to classify and recognize the fused features, so as to accurately determine the beat position of music. The parameter vector $w_2 \in R^{D_x}$ is extended into a parameter matrix $w_2 \in R^{r \times D_x}$, with r a hyperparameter indicating the count of attention weight vectors. Subsequently, the attention weight matrix $A = [a_1, \cdots, a_r]^T$ is derived, using the following calculation method:

$$A = soft\max\left(w_2 \phi\left(w_1 x^T\right)\right) \tag{10}$$

The aforementioned formula $soft\max \cdot$ is calculated along the second dimension of the input matrix to ensure each group of weight vectors sums to 1. The super parameter r, denoting the number of weight vectors, was set to 4 in the experiment. Using this attention weight matrix, a two-dimensional embedding matrix M is further computed to aggregate the music feature sequence:

$$M = AX \tag{11}$$

The i-th line vector m_i in M represents the weighted sum of feature vectors corresponding to each moment in the feature sequence. In the embedded music representation matrix M, each vector focuses on distinct parts of the entire feature sequence, enabling it to retain various musical features from the sequence.

5 EXPERIMENT AND ANALYSIS

This section will conduct experiments based on the widely used GTZAN data set, aiming at deeply analyzing the influence of different network structures on the performance of music beat detection and classification tasks, and comparing it with previous common algorithms in this field to evaluate the accuracy. To ensure experiment comprehensiveness and fairness, the data set is scientifically partitioned into training, verification, and test sets. The software and hardware configuration details are presented in Table 1.

Item	*Detailed Description*
Operating System	Windows 10
Programming Language	Python 3.8
Deep Learning Framework	TensorFlow 2.4
Data Processing Library	NumPy, Pandas
Processor	Intel Core i7
Memory	16GB RAM
Storage	500GB SSD
Graphics Processor	NVIDIA GeForce GTX 1080

Table 1: Experimental Software and Hardware Configuration.

During the experiment, a non-repeating random sampling technique was used to ensure that the dataset was divided into independent and non-intersecting training and validation sets. Among them, the training set contains 80000 data samples for training the network model; The validation set contains 10000 data samples for evaluating model performance and adjusting hyperparameters during the training process. After completing model training, use an independent test set to evaluate the final performance of the model. The beat detection error of the test set is shown in Figure 5, which intuitively shows the error distribution of the model when detecting beats. After calculation, the average beat detection error of the test set is ± 5 milliseconds.

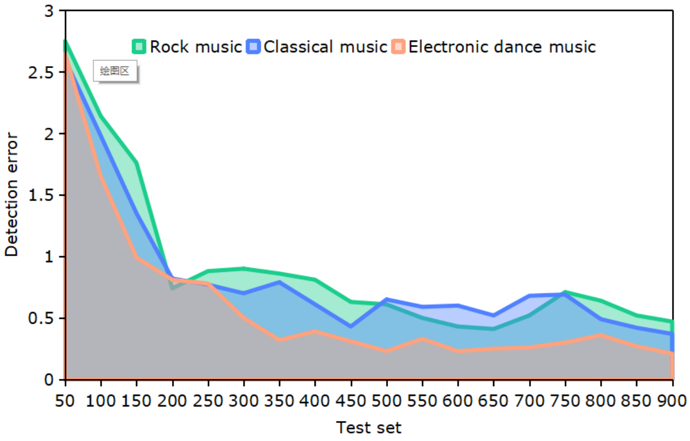

Figure 5: Beat detection error.

Figure 6 displays the beat detection accuracy of the test set, further validating the method's efficacy in this task. Notably, the method achieves high accuracy in detecting three music beat types: rock (96%), classical (94%), and electronic dance (97%), demonstrating its strong generalization and applicability.

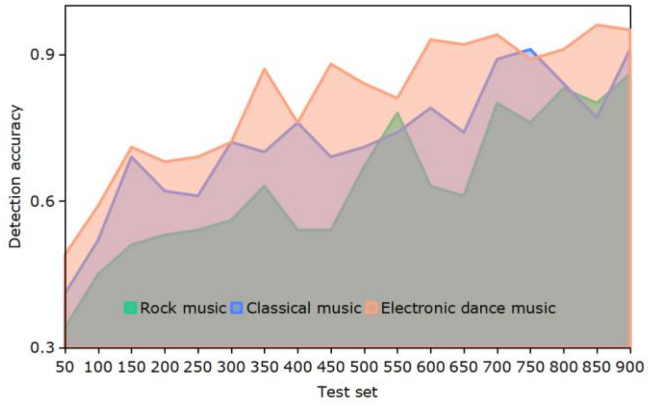

Figure 6: Accuracy of beat detection.

To comprehensively evaluate the model's training process, Figure 7 depicts the accuracy trends of the training and test sets across epochs. It illustrates the model's performance evolution and the disparity between training and test set accuracies, offering a solid foundation for further optimization. After 100 epochs, the model achieves stable accuracies of 98% on the training set and 95% on the test set, indicating good generalization and absence of overfitting.

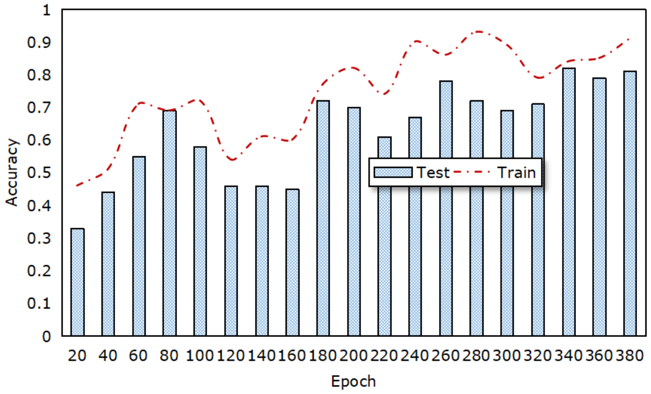

Figure 7: Accuracy of the training set and test set varies with epoch.

Finally, the beat detection time of the model is measured, as shown in Figure 8. The figure shows the beat detection time of the model when processing different music samples, which provides an important reference for evaluating the real-time performance and computational efficiency of the model. After measurement, the average beat detection time of each music sample is 20 milliseconds, which proves that the model has high computational efficiency and can meet the needs of real-time beat detection.

6 CONCLUSIONS

The main content of this article is to establish a neural network model for beat detection and related applications by extracting multimodal music features from music samples. Extract music features from both discrete models and continuous beat detection and fuse them into the constructed neural network model.

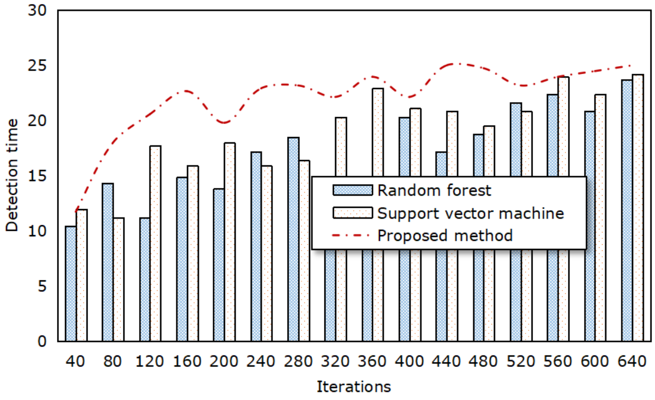

Figure 8: Beat detection time.

By improving the music beat detection model through three steps of preprocessing, feature extraction, and model optimization, the accuracy of recognition can be enhanced. At the end of the article, an intelligent music system was designed, which embedded beat detection and music generation modules. Treat music data as analogue signals. When processing this type of information, it is usually sampled periodically, but after quantization processing, it will bring a certain degree of quantization noise and distortion. So preprocessing is an indispensable part, and this article will adopt preprocessing methods such as pre-emphasis, windowing, and framing to make music information have short-term stationarity. The preprocessed sound wave information does not have emotional features, so the last step before sending the music samples to the classifier is to extract emotional features from the music samples. Merge two types of features in the high-dimensional space of WaveNet to preserve the original music features to the greatest extent possible. The preprocessed music data is more representative, laying the foundation for the subsequent recognition part. Finally, sentiment recognition was performed in the GAN network and compared horizontally with mainstream sentiment recognition models to obtain the final VA prediction regression value.

Jie Tian, https://orcid.org/0009-0004-6375-0332

REFERENCES

[1] Chaturvedi, V.; Kaur, A.-B.; Varshney, V.: Music mood and human emotion recognition based on physiological signals: a systematic review, Multimedia Systems, 28(1), 2022, 21-44. https://doi.org/10.1007/s00530-021-00786-6

[2] Coletta, A.; Marsico, M.; Panizzi, E.: MIMOSE: multimodal interaction for music orchestration sheet editors: An integrable multimodal music editor interaction system, Multimedia Tools and Applications, 78(23), 2019, 33041-33068. https://doi.org/10.1007/s11042-019-07838-0

[3] Correia, N.; Tanaka, A.: From GUI to AVUI: situating audiovisual user interfaces within human-computer interaction and related fields, EAI Endorsed Transactions on Creative Technologies, 8(27), 2021, 1-9. http://dx.doi.org/10.4108/eai.12-5-2021.169913

[4] Das, S.; Kolya, A.-K.: Detecting generic music features with single layer feedforward network using unsupervised Hebbian computation, International Journal of Distributed Artificial Intelligence (IJDAI), 12(2), 2020, 1-20. https://doi.org/10.4018/IJDAI.2020070101

[5] Das, S.; Satpathy, S.: Multimodal music mood classification framework for kokborok music, Solid State Technology, 63(6), 2021, 5320-5331. https://doi.org/10.1007/978-981-33-4299-6_14

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

[7] Forte, D.-L.: Music and discourse: A systemic-functional approach for music analysis in multimodal contexts, Multimodality & Society, 3(1), 2023, 69-81. https://doi.org/10.1177/26349795231153963

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

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

[10] Gorbunova, I.-B.; Plotnikov, K.-Y.: Music computer technologies in education as a tool for implementing the polymodality of musical perception, Musical Art and Education, 8(1), 2020, 25-40. https://doi.org/10.31862/2309-1428-2020-8-1-25-40

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

[12] Jandaghian, M.; Setayeshi, S.; Razzazi, F.: Music emotion recognition based on a modified brain emotional learning model, Multimedia Tools and Applications, 82(17), 2023, 26037-26061. https://doi.org/10.1007/s11042-023-14345-w

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

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

[15] Liao, Y.; Gui, Z.: An intelligent sparse feature extraction approach for music data component recognition and analysis of hybrid instruments, Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 45(5), 2023, 7785-7796. https://doi.org/10.3233/JIFS-231290

[16] Maba, A.: Computer-aided music education and musical creativity, Journal of Human Sciences, 2020, 17(3), 822-830. https://doi.org/10.14687/jhs.v17i3.5908

[17] Pandeya, Y.-R.; Lee, J.: Deep learning-based late fusion of multimodal information for emotion classification of music video, Multimedia Tools and Applications, 80(38), 2021, 1-19. https://doi.org/10.1007/s11042-020-08836-3

[18] Xu, L.; Zhang, S.: Music feature recognition and classification using a deep learning algorithm, International Journal of Computational Intelligence and Applications, 22(03), 2023, 2350012. https://doi.org/10.1142/S1469026823500128

[19] Yang, T.; Nazir, S.: A comprehensive overview of AI-enabled music classification and its influence in games, Soft Computing, 26(16), 2022, 7679-7693. https://doi.org/10.1007/s00500-022-06734-4