

Research on Deep Learning-Based Methods for Emotional Expression Recognition and Style Analysis in Piano Performance

Li Xiao

School of Preschool Education, Quanzhou Preschool Education College, Quanzhou 362000, Fujian, China

Corresponding author: Li Xiao (e-mail: xlsichu1069@163.com)

ABSTRACT Emotional expression recognition in piano performance requires fine-grained modeling of audio timbre, dynamic touch, rhythm fluctuation, and performance style. Existing methods based on single audio features or shallow statistics often fail to capture subtle emotional transitions and stylistic differences. This study proposes a deep-learning-based framework integrating multimodal feature fusion, temporal modeling, attention mechanisms, and emotion-style collaborative learning. Piano audio signals are converted into time-frequency representations through short-time Fourier transform and Mel-scale mapping, while MIDI control information encodes velocity, duration, and beat-offset features. Audio spectrogram features and MIDI control vectors are fused to construct a unified representation. A convolutional network extracts local time-frequency patterns related to accents and touch variations, and a temporal modeling module captures emotional evolution across performance sequences. An energy- and velocity-guided attention mechanism enhances key musical segments, including climaxes, strong beats, and ornaments. In addition, a style representation branch models rhythmic stability, dynamic fluctuation, and tempo evolution, while cross-task consistency constraints couple emotion recognition and style analysis. Experiments on paired audio-MIDI performance data show that the proposed method achieves 0.91 accuracy and 0.90 F1 score, improving by approximately 9% and 10% over unimodal baselines. The framework provides an audio signal processing and multimodal sequence-analysis method for intelligent music-performance evaluation.

INDEX TERMS audio signal processing, multimodal feature fusion, deep learning, temporal modeling, emotion recognition

I. INTRODUCTION

AS the music information processing and artificial intelligence technologies rapidly evolve, automatic identification of emotional expression in the piano performances has slowly become the significant line of research in the digital music analysis and intelligent music education. Emotional expression in piano performances is more dependent on multi-dimensional parameters like timbre variations, touch strength, and rhythm and its demonstration is more covert and multi-faceted than speech emotion recognition. Nevertheless, there are no techniques that predominantly focus on individual audio characteristics or other superficial statistical measures, and it is challenging to comprehensively describe the subtle emotional dynamics in performance. Moreover, although a few studies have proposed deep learning models, they nevertheless lack in dynamic modeling and representation of the difference in style, in particular, the systematic analysis of the intrinsic relationship between

the performance style and emotion, which makes them less effective in differentiating the similarity of the emotion categories, as well as in identifying complex expressions [1,2].

In a bid to solve the above problems, the paper suggests a deep learning-based solution to the problem of detecting and analyzing the emotional expression in piano performances. By fusing audio spectrograms and MIDI control information, a multimodal feature representation is constructed. Based on this, a convolutional neural network and a temporal modeling structure are combined to achieve joint modeling of local time-frequency features and the dynamic evolution of emotions. More so, a performance-based attention mechanism that operates on the basis of performance energy and dynamics information is proposed to boost the expressive power of the key performance segments. At the same time, a joint modeling emotion and style is developed to reach multi-task joint learning

in a common feature space and consistency constraints enhance the degree of semantic matching between the two. This approach constructively enhances the performance of models at various levels feature construction, emotion modeling, style expression and collaborative optimization, which offers a technical route with a high level of expressive capability and interpretability to identify emotions in piano performance.

II. LITERATURE REVIEW

Existing research has explored music information modeling and emotion recognition in multiple dimensions, but there are some differences in the methodological paths and research focuses. Li et al. jointly modeled structural information such as pitch and tempo in piano playing, and introduced Word2Vec and BiLSTM-CRF models to effectively improve the accuracy of sequence decision-making, which shows the important role of multi-feature fusion and sequence modeling in music tasks [3]. Furthermore, in the field of music generation, another study improved the generation quality and stability from the signal processing level by introducing MFCC features and optimizing the STFT processing process, and combined with the design of hybrid window functions to reduce spectral leakage [4], indicating that the underlying feature expression has a fundamental impact on the performance of music tasks. From the perspective of cognition and learning, Xu et al. revealed the moderate positive correlation between music learning and speech awareness through meta-analysis, and pointed out the moderating role of factors such as learning time [5], which provides theoretical support for the relationship between music emotion and cognitive ability. In the direction of emotion recognition, research has gradually developed from single features to multi-feature fusion and deep models. Wang and Huang achieved efficient classification of musical emotions by fusing MFCC and residual phase information and constructing an LSTM-BLS model, demonstrating the advantages of wide-depth structure in improving expressive ability and training efficiency [6]. Zhu et al. further combined multi-feature input with one-dimensional CNN structure, introduced attention weighting mechanism and incremental optimization strategy, effectively improving the generalization ability of the model on different datasets [7]. Overall, existing research has made some progress in feature fusion and model structure optimization, but most works still focus on single task modeling and lack a systematic characterization of the intrinsic relationship between performance style and emotional expression. This also provides an entry point for this paper to carry out research from the perspective of multimodal fusion and collaborative modeling.

III. RESEARCH METHODOLOGY DESIGN

Data Representation and Multimodal Feature Construction

To effectively depict the emotional information in piano performance, the audio signal is first modeled in the time

and frequency domains. The original audio is then segmented into frames using the Short-Time Fourier Transform (STFT), and its time-frequency representation is as follows:

$$X(\tau, \omega) = \sum_{n=-\infty}^{+\infty} x(n)w(n - \tau)e^{-j\omega n} \quad (1)$$

Based on this, a Mel frequency scale is introduced to perform a nonlinear mapping of the spectrum:

$$M(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (2)$$

This allows us to extract spectral energy distribution characteristics that better match the perception characteristics of the human ear.

At the same time, the performance control information is structured and encoded using MIDI data, representing parameters such as velocity, duration, and beat offset as vectors:

$$C = [v_i, d_i, \Delta t_i] \quad (3)$$

This is to characterize the dynamic control behavior during the performance process. Furthermore, audio features and control features are weighted and fused to construct a unified multimodal representation:

$$F = \alpha F_a \oplus \beta F_m \quad (4)$$

Here, α and β are weighting coefficients. This method enables the collaborative expression of information from different modalities, improving the discriminative power of emotional features while ensuring the compactness of the representation.

Construction of Emotion Expression Recognition Model Local Feature Learning Model Based on Convolutional Neural Network

To address the impact of tonal details and touch variations on emotional expression in piano performance, a convolutional feature extraction model based on the spectrogram is constructed. Using the Mel spectrogram ($S(t, f)$) as input, local performance features are extracted along the time-frequency dimension using convolutional kernels, expressed as follows:

$$H_k(t, f) = \sigma \left(\sum_{\Delta t, \Delta f} W_k(\Delta t, \Delta f) \cdot S(t + \Delta t, f + \Delta f) \right) \quad (5)$$

Among them, the convolution kernel W_k focuses on micro-patterns related to emotion, such as spectral energy mutations (corresponding to accents) and high-frequency attenuation (corresponding to soft touch), thereby achieving local modeling of the subtle expression of the performance. *Dynamic Sentiment Modeling Method Based on Temporal Networks*

Considering the significant temporal progression of emotions in piano performance, a temporal modeling mechanism based on performance sequences is introduced. The convolutional output sequence is represented as follows (x_t), and the emotional state update process is constructed by combining changes in playing speed and dynamics:

$$h_t = \tanh(W_1 x_t + W_2 h_{t-1} + W_3 \Delta v_t) \quad (6)$$

Here, Δv_t represents the change in intensity between adjacent moments (reflecting emotional tension). This modeling method can capture the influence of performance behaviors such as crescendo, diminuendo, and rhythmic stretching on the continuous evolution of emotion, realizing a mapping from “note sequence” to “emotional trajectory”. *A sentiment discrimination model incorporating attention mechanisms*

To highlight the dominant role of key performance passages (such as climaxes, strong beats, or ornaments) in emotional expression, an attention weighting mechanism oriented towards musical structure is introduced. Attention weights are calculated based on spectrogram energy and performance control information.

$$\alpha_t = \frac{\exp(\gamma E_t + \delta v_t)}{\sum_k \exp(\gamma E_k + \delta v_k)} \quad (7)$$

Where E_t represents the spectral energy at that moment, v_t and is the corresponding intensity parameter. Further, the emotion aggregation representation is obtained:

$$s = \sum_t \alpha_t h_t \quad (8)$$

This mechanism enables the model to adaptively focus on the performance segments where emotional expression is most pronounced, thereby improving the accuracy of distinguishing complex emotions, such as lyrical versus dramatic.

IV. MODELING AND ANALYSIS METHODS FOR PERFORMANCE STYLE CHARACTERISTICS

Parametric Representation Model of Style Features

To address the differences in rhythm control, touch intensity, and tempo handling among piano playing styles, a parametric feature representation oriented towards performance behavior is constructed. First, rhythmic stability is characterized by fluctuations in beat intervals, expressed as:

$$R = \frac{1}{T} \sum_{t=1}^T |\Delta t_t - \overline{\Delta t}| \quad (9)$$

Here, Δt_t represents the time interval between adjacent notes, reflecting the evenness of the playing rhythm. Secondly, velocity distribution characteristics are constructed based on MIDI velocity:

$$D = \frac{1}{N} \sum_{i=1}^N (v_i - \bar{v})^2 \quad (10)$$

It is used to characterize the dynamic fluctuations and emotional tension in a performance. Furthermore, it depicts the trend of tempo evolution through changes in the tempo of adjacent notes:

$$V_t = \frac{d_t - d_{t-1}}{d_{t-1}} \quad (11)$$

By synthesizing stylistic feature vectors, a multi-dimensional characterization of performance style can be achieved.

Style Representation Learning Method Based on Deep Embedding

To enhance the abstract expressive power of style features, an embedding mapping mechanism oriented towards style representation is constructed based on parameterized features. The original style vector (x) is projected into a low-dimensional space through a nonlinear mapping function:

$$z = \phi(Wx + b) \quad (12)$$

The mapping process primarily preserves stylistic factors such as rhythmic fluctuations and dynamic variations. To enhance the differentiation between different performance styles, a stylistic contrast constraint is introduced, causing similar styles to cluster more closely within the embedded space while keeping dissimilar styles separate. The constraint takes the following form:

$$L = \max(0, d(z_i, z_j) - d(z_i, z_k) + m) \quad (13)$$

Here, (z_i, z_j) represents similar style performances, and (z_k) represents samples with different styles.

Style Similarity Measurement and Cluster Analysis Methods

After obtaining the style embedding representation, a measurement mechanism oriented towards performance style differences is further constructed. Style similarity is defined based on the directional consistency between embedding vectors:

$$\text{Sim}(i, j) = \frac{z_i \cdot z_j}{\|z_i\| \|z_j\|} \quad (14)$$

This indicator reflects the overall similarity in rhythm control and dynamic expression across different performances. Based on this, performance samples are aggregated according to the distribution of stylistic features, and style segmentation is achieved by minimizing intra-class differences.

$$J = \sum_k \sum_{i \in C_k} |z_i - \mu_k|^2 \quad (15)$$

Here, μ_k represents the style center of type (k).

V. COLLABORATIVE MODELING MECHANISM OF EMOTION AND STYLE

Affect-Style Joint Representation Learning Framework

To characterize the intrinsic relationship of “style-driven emotional expression” in piano performance, a joint learning framework for emotion recognition and style representation is constructed. In a shared feature space, multimodal input features (F) are mapped to a unified representation:

$$h = \psi(F) \quad (16)$$

Based on this, establish emotion branches and style branches respectively:

$$y_e = \phi_e(h), \quad y_s = \phi_s(h) \quad (17)$$

Here, y_e represents the sentiment category output, and y_s represents the style embedding representation.

Cross-task constraints and feature sharing mechanism

To strengthen the coupling between sentiment and style, a cross-task constraint mechanism is introduced on top of the shared encoding layer. By constraining the consistency of sentiment and style features, the two types of representations maintain relevance in the semantic space. The constraint takes the form of:

$$L_c = \|g(y_e) - y_s\|^2 \quad (18)$$

Here, $g(\cdot)$ represents the mapping function from emotion to style. This mechanism avoids the information fragmentation problem caused by independent learning of two tasks, allowing the model to retain stylistic structural information while learning emotion categories, thus more accurately reflecting the inherent consistency of performance expression.

Expression Consistency Optimization Strategy

Furthermore, to ensure the semantic matching relationship between sentiment recognition results and style features, an expression consistency optimization strategy is introduced. Sentiment classification loss, style representation constraints, and consistency loss are integrated into the overall loss function:

$$L = \lambda_1 L_e + \lambda_2 L_s + \lambda_3 L_c \quad (19)$$

Where L_e is the emotion recognition error, L_s is the style representation constraint term, L_c and is the consistency loss.

VI. EXPERIMENTAL PERFORMANCE EVALUATION

Dataset Construction and Annotation Methods

The experimental data comprised piano audio and MIDI files across lyrical, dramatic, and balanced styles. While the core dataset utilized high-fidelity recordings, we introduced a noise-augmented validation set containing environmental interference and performer-specific timing idiosyncrasies. This allows for a logical assessment of the model’s robustness when processing non-standardized MIDI inputs and varying acoustic qualities typical of live performance

environments. The audio was segmented into discrete samples according to the musical phrase structure. Emotional labels were initially categorized using primary discrete states (e.g., joy, sadness, tension); however, to account for the transitional and complex nature of musical expression (such as bittersweet or mixed emotional states), a soft-labeling approach was integrated. This allows the model to capture the semantic overlap between categories and reflects the logical progression of emotions often masked by rigid classification. Multi-annotator cross-validation mechanism was implemented to provide label consistency and a consistency coefficient was used to select valid samples. The experimental dataset consists of 1,200 paired audio and MIDI recordings totaling approximately 45 hours of performance data. These samples feature 30 professional performers across lyrical, dramatic, and balanced styles. The audio was segmented into 8,500 distinct clips based on musical phrase structures. For experimental validation, the data was partitioned into training, validation, and testing sets using a 70:15:15 ratio, ensuring that recordings from the same performer did not overlap between the training and testing phases. A consistency coefficient was applied to ensure high-quality labeling among expert annotators.

Experimental Setup and Comparison Methods

In the scenario of the equal input of the features, several sets of comparative models were developed to isolate the specific contributions of both data modality and architectural components. These included high-complexity unimodal benchmarks (incorporating identical attention and temporal layers to those in the proposed model), as well as a multimodal ablation variant lacking the collaborative mechanism. This ensures that the observed 9-10% performance gain is logically attributable to the synergistic fusion of audio and MIDI data rather than mere increases in network depth. Each model had the same number of training epochs and batch size, the learning rate, and the parameter initiation strategy were maintained during the optimization to guarantee the comparability of the results.

Analysis of Emotion Recognition Results

Model performance was evaluated using classification accuracy and F1 score, supplemented by 95% confidence intervals and p-value calculations ($p < 0.05$) to verify statistical significance. While the overall accuracy reached 0.91 and the F1 score was 0.90, per-class analysis revealed that ‘Joy’ and ‘Tension’ achieved F1 scores above 0.92, whereas the ‘Sadness’ category showed slightly lower precision due to its subtle overlap with lyrical styles. This granular evaluation confirms that the multimodal approach provides a statistically robust improvement over baseline models. Further analysis revealed that incorporating an attention mechanism makes the model more sensitive to key performance segments, thereby reducing misclassifications.

As Figure 1 shows, the various models have a large gap in performance of the emotion recognition task. Of the unimodal models, the audio-based CNN model has an accuracy of 0.82, F1 of 0.80, whereas the LSTM model using only

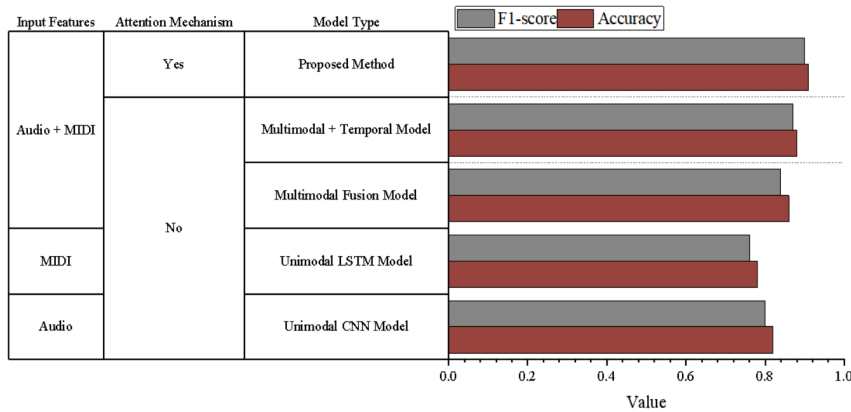


FIGURE 1. Performance comparison results of emotion recognition models

MIDI features has 0.78 and 0.76 respectively, suggesting that one modality alone is not enough to capture all the emotional information of piano performance. Introducing multimodal fusion can greatly enhance the performance of the models, both the accuracy and the F1 score rise to 0.86 and 0.84 respectively, indicating the complementary nature of information of audio and the performance control when trying to express emotion. Moreover, using temporal modeling further improves the performance of the model to 0.88 and 0.87 which confirms the goodness of the dynamic modeling in the process of emotion evolution. Lastly, the suggested approach yields the best performance, with the addition of an attention mechanism, an accuracy of 0.91 and an F1 score of 0.90, which is a 9 per cent and 10 per cent improvement over the original unimodal model. This finding shows that the attention mechanism is capable of effective focusing on the important performance segments, minimize the misclassification of similar emotion categories and therefore greatly enhanced the overall performance of recognition.

Ablation Experiment Analysis

Removing the multimodal fusion, temporal modeling, and attention mechanisms in a progressive manner, the contribution of each of the components to the overall performance was examined. The findings indicate that the convolutional architecture mainly boosts the ability to represent local features, the temporal module boosts the dynamic modeling, and the attention mechanism is essential in the recognition of complex emotions. Eliminating the collaborative modeling mechanism results in a considerable degradation of performance, which also confirms its central position in the overall framework.

TABLE 1. Ablation Experiment Results

Model Configuration	Multi-modal fusion	Temporal modeling	Attention mechanism	Collaborative modeling	Accuracy	F1 score
Complete Model (The Method Used in This Paper)	✓	✓	✓	✓	0.91	0.90
Remove collaborative modeling	✓	✓	✓	×	0.88	0.87
Remove attention mechanism	✓	✓	×	✓	0.87	0.86
Remove timing modeling	✓	×	✓	✓	0.85	0.84
Remove multimodal fusion	×	✓	✓	✓	0.83	0.81
Only retain the CNN (base model)	×	×	×	×	0.80	0.78

Table 1 indicates that the various modules greatly affect the model performance. The overall model gives the highest scores on all measures, where the accuracy is 0.91 and the F1 score is 0.90, meaning that the synergistic effect of each of the modules is a successful way to increase emotion recognition. The accuracy and F1 score drop to 0.88 and 0.87 respectively when the collaborative modeling mechanism is dropped, indicating that the factor of emotion and style joint modeling is a significant supporting factor to the overall performance.

Additional artificially induced removal of attention mechanism results in a further drop in model performance (0.87/0.86) which implies that the weighted modeling of key performance segments has a considerable impact on the ability to distinguish fine-grained emotions. The accuracy and F1 score reduced to 0.85 and 0.84, respectively, when temporal modeling module was removed, which shows that the use of the static features alone is not enough to describe the dynamics of the emotions changing over time. By contrast, the drop in performance was even more prominent when multimodal fusion had been eliminated (0.83/0.81), indicating that audio and MIDI features were highly complementary with each other in the emotion expression modeling task and played a vital role as a starting point to enhancing recognition performance. The lowest performance was achieved in case of sole basic CNN model retained (0.80/0.78) and again confirms the drawbacks of a single structure in complex emotion recognition exercises.

VII. CONCLUSION

The paper deals with the issue of identification of emotional expressions during piano performances. It develops an analysis approach that combines deep learning model with multimodal characteristics. It successfully captures emotional variation in performance by integrating convolutional networks, time series encoding, and attention to understand it. At the same time, the mechanism of emotion-style co-modeling is proposed to provide the joint expression of the two kinds of information in a common feature space, which greatly enhances the ability to discriminate and

stability of the model in complicated emotion recognition tasks. The effectiveness of the multimodal fusion and co-modeling strategy is confirmed by experimental results that indicate that the proposed method is more accurate and has a higher F1 score as compared to the comparative models. It is important to note that this approach continues to use rather idealized labeled data and standardized performance samples. Its generalization capability should be enhanced in the presence of noise interference and stylistic heterogeneity in the actual performance settings. The future studies may be directed at the work on the expansion of cross-performer data, self-supervised learning, and finer-grained emotion modeling to additionally increase the adaptability and value of the model in practice.

AUTHOR CONTRIBUTIONS

Li Xiao designed, collected and analyzed the data, and drafted the manuscript. Li Xiao conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Li Xiao participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] Y. Zhang, Y. Feng, and Q. Ding, "A Multi-kernel Convolutional Neural Network for Multi-emotion Classification of Music Lyrics Based on Word2Vec Word Embedding," *Science Technology and Engineering*, vol. 24, no. 20, pp. 8598-8605, 2024.
- [2] C. Xiao, "A Brief Analysis of the Path to Improve the Literacy and Ability of Piano Art Instructors," *Modern Education Frontiers*, vol. 6, no. 9, pp. 15-17, 2025.
- [3] Q. Li, Z. Wu, and X. Guan, "A method for generating piano fingerings by combining deep musical score features," *Journal of Intelligent Systems*, vol. 18, no. 6, pp. 1287-1294, 2023.
- [4] Y. Li, L. Wang, Y. Xue, et al., "Analysis and research on intelligent music generation system based on short-time Fourier transform," *Journal of Intelligent Systems*, vol. 20, no. 3, pp. 750-760, 2025.
- [5] Y. Xu, M. Cheng, H. Shao, et al., "The relationship between music learning and phonological awareness: evidence from metaanalysis," *Psychological Technology and Application*, vol. 11, no. 12, pp. 705-721, 2023.
- [6] J. Wang and R. Huang, "Music Emotion Recognition Based on Wide-Depth Learning Network," *Journal of East China University of Science and Technology (Natural Science Edition)*, vol. 48, no. 3, pp. 373-380, 2022.
- [7] Y. Zhu, T. Feng, M. Zhang, et al., "Convolutional speech emotion recognition network based on incremental method," *Journal of Shanghai University (Natural Science Edition)*, vol. 29, no. 1, pp. 24-40, 2023.