International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 6373-6382

https://doi.org/10.70517/ijhsa464541

A Constructivist-Inspired Deep Learning Framework for Enhanced Musical Theatre Singing Analysis and Signal Separation

Linlong Jiang[1,*]
[1] School of music, University of Sanya, Sanya, Hainan, 572000, China
Corresponding authors: (e-mail: 18907762163@163.com).

Abstract Musical theatre performance integrates emotional expression, character construction, and dramatic development, where singing plays a pivotal role in bridging narrative and music. However, traditional approaches to musical singing and analysis often overlook the contextual and structural nuances embedded in scripts and scores. This study proposes a constructivist-inspired learning and signal processing framework that enhances the accuracy and interpretability of musical theatre singing through deep neural collaborative filtering. Leveraging spectrogram analysis, encoder-decoder architectures, and SA attention-based feature extraction, we construct a multi-module system to improve the fidelity of vocal signal separation and the interpretive quality of performance modeling. Empirical results demonstrate significant gains in sub-module construction accuracy and signal restoration performance, offering a robust technical foundation for intelligent musical analysis.

Index Terms Neural Collaborative Filtering, Algorithm Optimization, Dynamic Performance, Musical Theatre Performance

I. Introduction

Musical theatre, as a composite art form, uniquely blends vocal performance, dramatic narrative, and stage expression to convey character emotion and promote narrative development. Among its core elements, singing under prescribed dramatic situations serves as both an expressive and structural mechanism that enhances plot progression and character depth [1]. However, in current vocal training and musical theatre pedagogy, many students and performers prioritize technical vocal execution while neglecting the dramaturgical context embedded in scripts and scores. This imbalance often results in disjointed performances where musical interpretation lacks emotional credibility and narrative cohesion [2]. Therefore, integrating contextual analysis with vocal instruction has become essential in enhancing the interpretive richness of musical theatre performance.

Within the broader discourse of education, constructivist theory offers a compelling pedagogical foundation for addressing this challenge. Constructivism emphasizes knowledge construction through experience, active participation, and contextual understanding, aligning well with the demands of musical theatre training [3]. Its principles—namely the constructive nature of knowledge and cognition—have profoundly influenced modern curriculum reform, including in the arts and music education domains [4], [5]. When applied to musical theatre, constructivist learning encourages performers to actively interpret roles, internalize emotional logic, and engage in experiential rehearsal, thereby enabling a more integrated approach to musical storytelling.

Despite these advancements, scientific inquiry into musical theatre singing remains limited, particularly in the context of signal processing, music information retrieval, and intelligent performance analysis. Prior research in music signal separation has largely focused on algorithmic design and spectral modeling, often constrained by rigid assumptions that fail in complex, multi-source scenarios [6]-[8]. Conventional methods rely on handcrafted features or linear modeling techniques, which can be insufficient when handling overlapping vocals and accompaniment. Furthermore, phase information, a crucial determinant of signal fidelity, is frequently neglected or inaccurately estimated, leading to degraded signal reconstruction [9].

To address these technical limitations, recent developments in deep learning—particularly encoder-decoder networks, convolutional structures, and attention mechanisms—have shown strong potential in learning high-level representations of music source signals. Time-frequency analysis combined with deep neural networks (DNNs), spectral attention (SA) modules, and feature extraction matrices (FEM) allows for accurate signal modeling and effective separation of musical components. These technologies not only improve signal reconstruction but also provide analytical insights into vocal dynamics and interpretive characteristics [10]-[12].

This study builds upon interdisciplinary insights from performance theory, education, and artificial intelligence to propose a novel framework for musical theatre singing analysis. Grounded in constructivist learning principles, the framework incorporates deep neural collaborative filtering, SA-based encoder-decoder architectures, and spectral decomposition techniques to enhance both artistic interpretation and music signal fidelity. Empirical experiments validate the framework's effectiveness in improving module construction precision, phase-amplitude spectrum modeling, and the interpretive quality of musical performance.

II. Improved neural collaborative filtering

If move the speaker diaphragm according to the recorded waveform, the sound is reproduced. A multi-channel signal simply consists of several waveforms captured by multiple microphones. Typically, a music signal is stereo, a combination of multiple channel signals. In general, an audio signal can be represented by time as an independent variable. In the field of signal processing, this one-dimensional audio signal is usually analyzed by Fourier transform. With the help of FT analysis, the distribution of different frequency components contained in the audio signal and their signal amplitudes. For stationary signals, there are two variables to focus on. They are time and frequency. With the help of FT, we can analyze the time domain characteristics of the signal in the time domain, and we can also convert it to the frequency domain to analyze the frequency domain characteristics. In particular, by performing FT analysis on the music signal, we can not only obtain the representation of the music signal in the frequency domain, but also obtain the energy distribution of the music signal. However, the Fourier transform analyzes a whole music signal, which has limitations. When we want to analyze the relationship between the frequency domain characteristic information of the signal and the time variable, we can no longer use FT to analyze it, and we can only analyze it separately in the time domain. and research in the frequency domain. FT cannot analyze the time-dependent relationship of frequency information in the signal, but it can analyze the relationship between the frequency components and phases contained in the entire signal and can analyze the amplitude distribution of different frequency components. In order to analyze the relationship between the frequency characteristics of the signal and the variable time, the time-frequency analysis method can be used. Its calculation method is shown in formula (1):

$$STFT(f,t)=\sum_{-\infty}^{+\infty}\left[x(t)m(t-\tau)\right]e^{-j2\pi ft}dt \tag{1}$$

Selecting a suitable window function when using $STFT$ analysis will have a greater impact on the analysis results. If the width of the window function is narrow, the frequency resolution of the signal will be low. Conversely, if the width of the window function is wide, the time resolution of the signal will be low. Figure 1 is an example of a human voice amplitude spectrogram of a music signal.

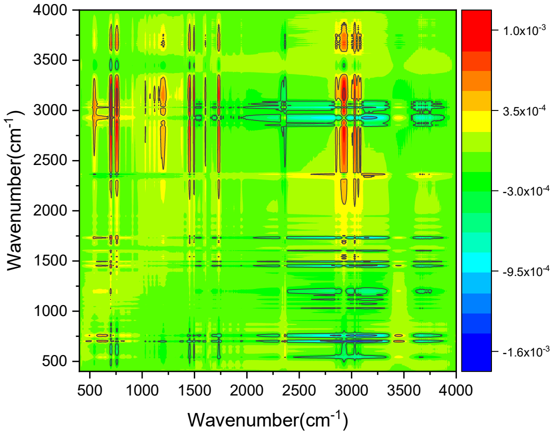


Figure 1: Example of vocal amplitude spectrogram of music signal

II. A.Deep Neural Network Collaboration and Convolution Operation

The fully connected neural network needs to calculate the weight information of the neurons in the whole network, which leads to the large amount of parameters of this kind of network, which also reduces the training speed of the entire model and brings about the problem of overfitting of the model. And it can effectively solve the problems encountered when the fully connected neural network processes image feature information.

The process of convolution is actually the process of weighted summation. The number of convolution kernels used in this convolution operation corresponds to the number of output channels of the convolution operation. The schematic Figure 2 of its convolution operation is as follows:

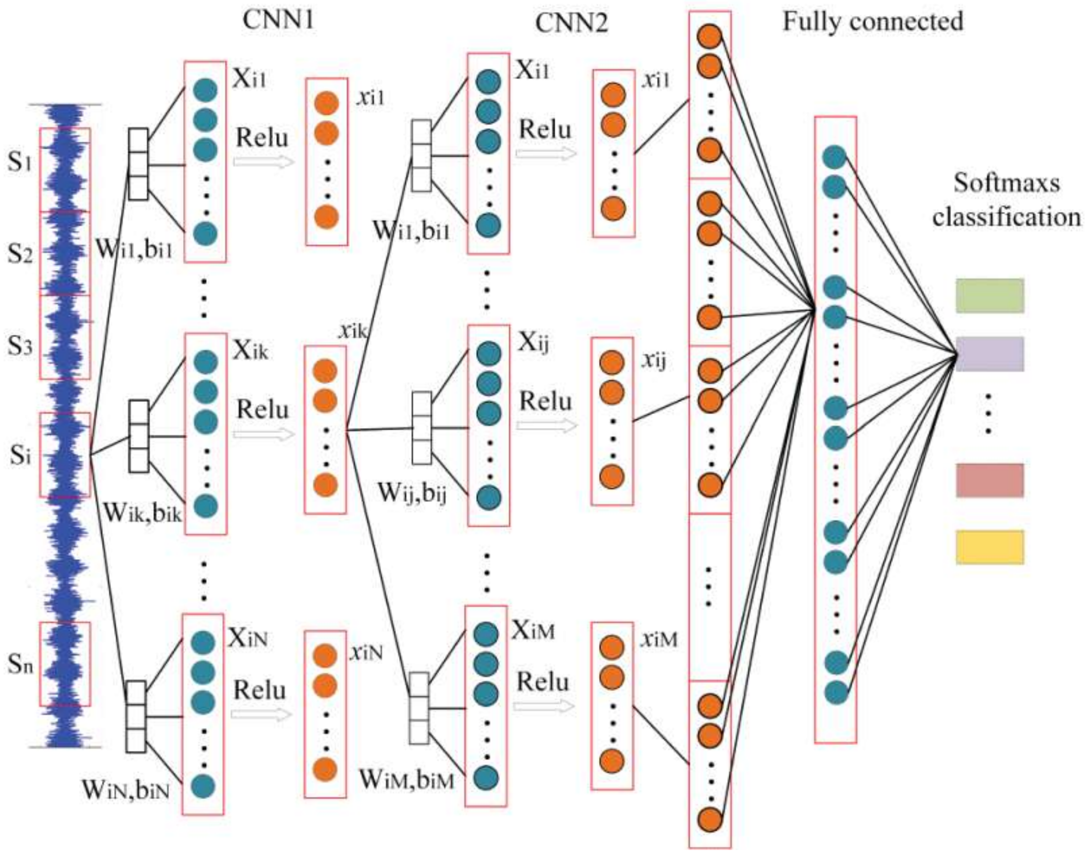


Figure 2: Schematic diagram of neural network convolution operation

Then move the window according to the preset fixed step size until all the input data are taken, and then the new eigenvalues will be spliced in order. The pooling layer directly reduces the resolution of the input features, reduces the calculation amount of the entire model, and also allows the network to obtain spatial invariance, which is also the extraction of features again. Figure 3 is a schematic diagram of the pooling operation.

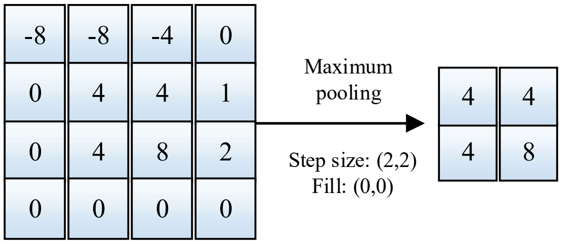


Figure 3: Schematic diagram of pooling operation

II. B.Normalization

Each time the data passes through a network layer and then is output, the parameters of the entire network layer will be updated. As the data is transmitted in each layer of the network, the distribution characteristics of the data must also be different from when it was first input to the network. If it is not corrected, The network needs to adapt to the input data of different distributions, and the learning cost of the entire network will be greatly increased, resulting in a slow gradient descent of the network and a long training time. This is the "Internal Covariate Shift" problem that has plagued researchers for a long time. In order to solve the problem, researchers have proposed many effective normalization processing algorithms to adjust the data distribution of each layer of the network.

Therefore, the most important thing for user-based collaborative filtering recommendation is to find the K users that are most "similar" to the target user. The similarity calculation formula is as follows:

$$sim(x,y)=\cos(x,y)=\frac{x,y}{\|x\|.\|y\|}=\frac{\sum_{k\in I^{r}xk^{r}yk}}{\sqrt{\sum_{k\in I_x{}^{r^2}yk}}\sqrt{\sum_{k\in I_x{}^{r^2}yk}}} \tag{2}$$

Taking into account the difference in evaluation scales among different users, a modified cosine similarity is proposed, and its formula is as follows:

$$sim(x,y)=\frac{\sum_{k\in I_{xy}}\left(r_{xy}-\overline{r_x}\right)\sum_{k\in I_{xy}}\left(r_{xy}-\overline{r_x}\right)}{\sqrt{\sum_{k\in I_x}\left(rxy-\overline{r_x}\right)^2}\sqrt{\sum_{k\in I_x}\left(rxy-\overline{r_x}\right)^2}} \tag{3}$$

Therefore, the current recommendation algorithms are all hybrid recommendation models, but it is still necessary to know the applicable scenarios, advantages and disadvantages of each recommendation model, so as to distinguish the priority and improve the model in the future recommendation tasks. Table 1 is a summary of common problems of several commonly used recommendation algorithms.

Table 1: Summary of common problems of several commonly used recommendation algorithms

Recommended algorithm	Whether to alleviate data sparsity	Whether to alleviate cold start	Generate personalized recommendation
User-based CF	NO	NO	NO
Item-based CF	NO	NO	YES
LFM	YES	NO	YES
Content -based	YES	YES	YES
Paper model	YES	YES	YES

The Skip-Gram model predicts the probability of occurrence of surrounding context words through several central words in a word sequence, and then smoothly predicts the occurrence probability of all words in the word sequence, as shown in Figure 4:

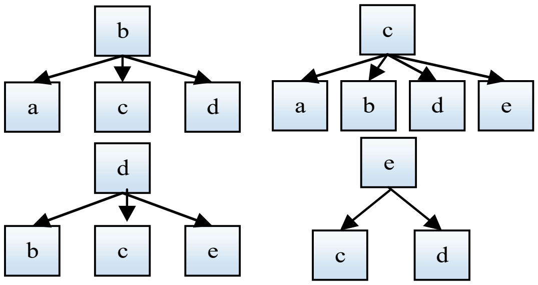
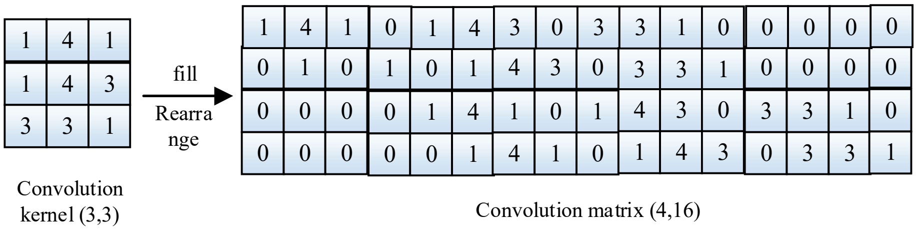


Figure 4: Schematic diagram of skip gram model4 Methods

In musicals, if the music signal is to be transmitted without distortion over long distances, it is necessary to encode the signal to be transmitted, convert the original signal into another encoded signal that can be easily transmitted for transmission, and then re-encode the signal at the receiving end of the signal. The signal is decoded to obtain the original signal. The encoder-decoder network structure is a network structure that applies encoding and decoding technology to the field of deep learning. The encoder in this network structure corresponds to the encoder, which can effectively convert an input data into Other types of feature data are output, but the main feature information in the original data is not lost, so as to facilitate the operations we need on the feature data. The decoder part, as the decoder, is responsible for restoring the data features extracted by the former encoding to the scale and dimension of the original input data. In deep learning, you can use all convolutional layers to build an encoder-decoder network, you can also use RNN to build the encoder and decoder in the network, or you can mix and match, use CNN to build the encoder of the network, and use RNN or LSTM to build The decoder of the network to meet the processing needs of the entire network for feature informationbeg. The schematic diagram of its convolution matrix is shown in Figure 5:



1	4	1
1	4	3
3	3	1

1	4	1	0	1	4	3	0	3	3	1	0	0	0	0	0
0	1	0	1	0	1	4	3	0	3	3	1	0	0	0	0
0	0	0	0	1	4	1	0	1	4	3	0	3	3	1	0
0	0	0	0	0	1	4	1	0	1	4	3	0	3	3	1

Figure 5: Experimental simulation diagram of convolution matrix

For audio signals, each harmonic component contained in the reconstructed signal is inseparable from the amplitude spectrum and phase spectrum corresponding to each harmonic component, and once the phase spectrum is disturbed, the reconstructed time domain signal will have a very large impact. , so the phase spectrum is not manipulated in this chapter. The frame diagram of its music source feature extraction and separation algorithm is shown in Figure 6.

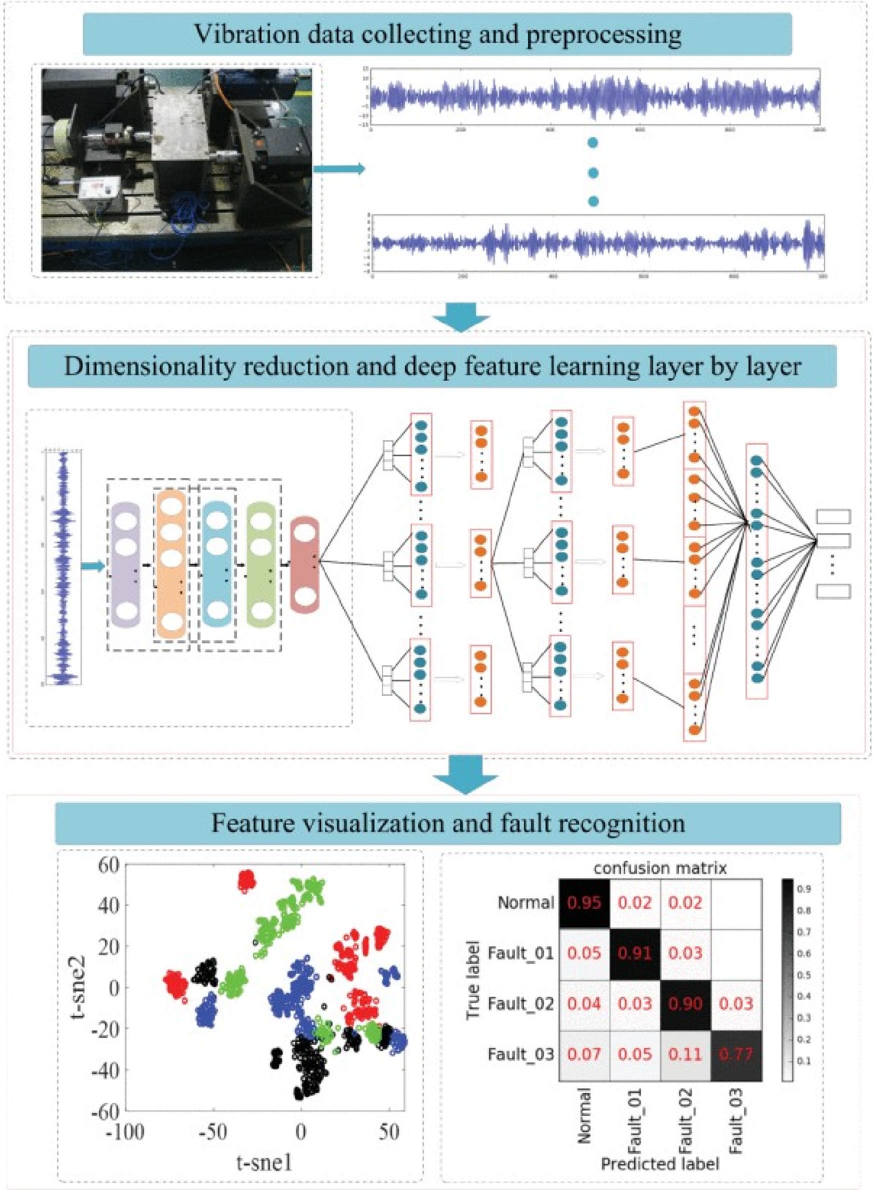
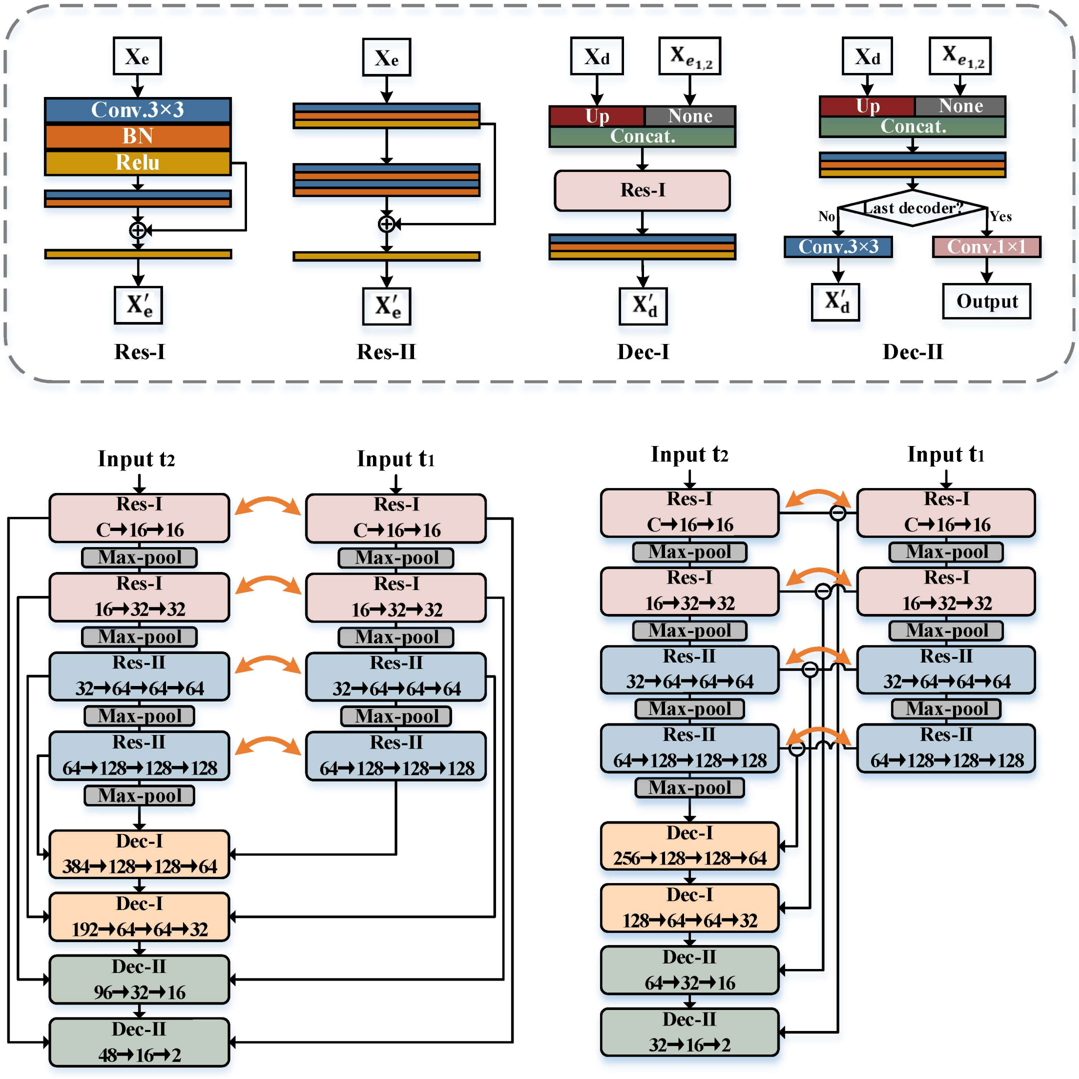


Figure 6: framework diagram of music source feature extraction and separation algorithm

The structure of the deep convolutional encoder-decoder network model based on SA attention mechanism proposed in this paper is shown in Figure 7.



(a) (b)

Figure 7: SA-based Encoder-decoder network

Among them, each up sampling block consists of 5 network layers, which are bilinear interpolation layer (BI), transposed convolution layer of size (3, 3), BN layer, dropout layer and Rely activation layer. The use of transposed convolution to construct up sampling blocks is abandoned, and bilinear interpolation is used to achieve up sampling, which reduces the amount of parameters and achieves the purpose of up sampling the feature map. The experimental environment is shown in Table 2.

Table 2: Experimental environment and configuration table

Environment name	Specific configuration
Operating system	Ubuntu 18. 04
Development language	Python3
Deep learning framework	Pytorch1.8
Integrated development environment	Vscode
CPU	15-10400f
CPU	GTX1080
Memory	32G
Hard disk	500G

III. Case study

III. A. *Improve the construction accuracy of the musical sub-module*

For musicals, the operation of aggregating the feature maps of the same level in the FEM structure is obtained by direct addition. The feature map of the input FEM is restored to the dimension before the input domain FEM module after the last layer of deconvolution upsampling module, and then the feature information after FEM feature extraction is obtained after passing through SAM. FEM does not change the size and number of channels of the input feature map. It uses a more efficient structure to extract the feature information of the input data. At the same time, FEM can be easily embedded into the convolutional neural network, which will increase a certain amount of parameters. , but can effectively increase the feature extraction capability of the entire network. The schematic diagram of its accuracy is shown in Figure 8.

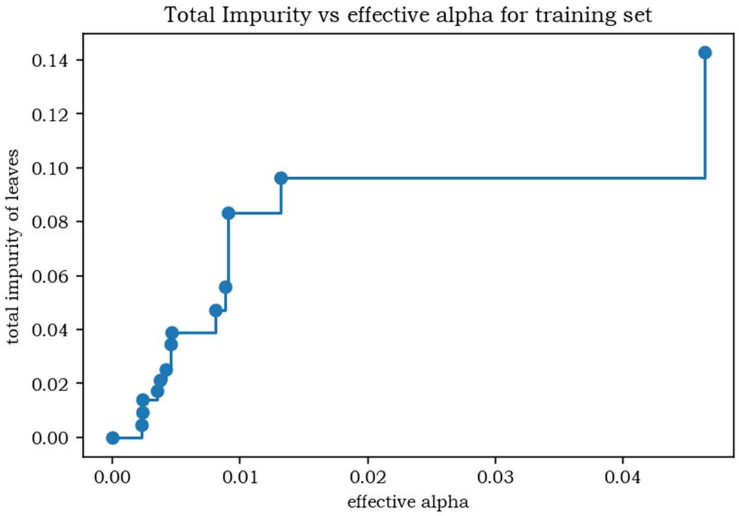


Figure 8: Schematic diagram of sub-module accuracy

Use two DNNs to extract features from the amplitude spectrum and use DNN to extract features from the phase spectrum of the mixed music signal. The amplitude spectrum feature is added to the network, and finally the two networks are used to predict the amplitude spectrum and phase spectrum of the music source signal, and then the separated music source signal can be reconstructed in the time domain with the help of ISTFT.

III. B. *Improve the effect of music interpretation*

After the partial derivative of the original phase spectrum is calculated in time and frequency, the phase compensation is performed, and then normalized to obtain two equal magnitudes. The modified phase spectrum is obtained by splicing the two phase spectrums in the channel direction and then inputting them into the FEM module proposed in Chapter 3, and then inputting the output of the FEM and the amplitude spectrum into the DNN. Perform feature information fusion and extraction to obtain the amplitude spectrum of the music source signal predicted by DNN, and then perform ISTFT transformation together with the original phase spectrum of the mixed music signal to obtain the separated and predicted time domain music source signal. As shown in Figure 9.

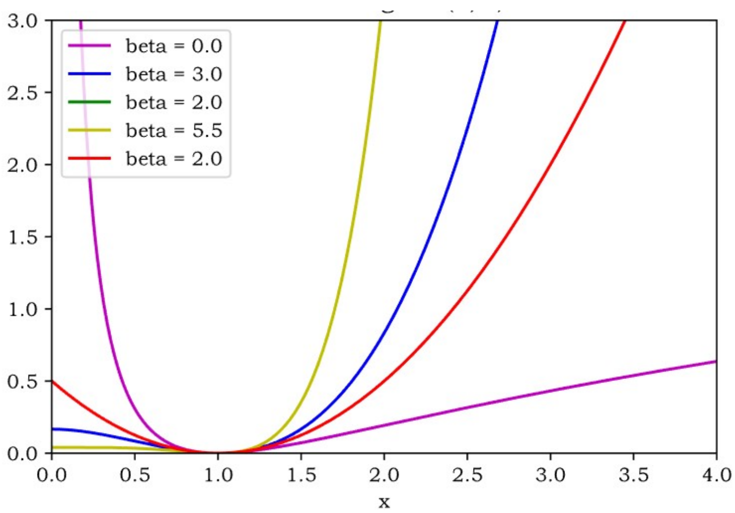


Figure 9: Amplitude spectrum and phase spectrum of music source signal

The Figure 10 illustrates the comparative effects of four types of contextual input—No Context, Text Prompt, Script Performance, and Mirror Imitation—on musical singing performance across three dimensions: emotional expressiveness, articulation clarity, and narrative advancement. The data reveal that situational guidance significantly influences performance outcomes. Notably, the "Script Performance" condition resulted in the highest scores across all dimensions, indicating that embedding the performer within a dramatic context enhances their expressive and narrative delivery. This suggests that situational immersion fosters more authentic and effective communication of character and plot. Conversely, the "No Context" group yielded the lowest scores, affirming the limitations of purely technical rehearsal without interpretive grounding. These findings support the pedagogical value of constructivist, experience-based training in musical education.

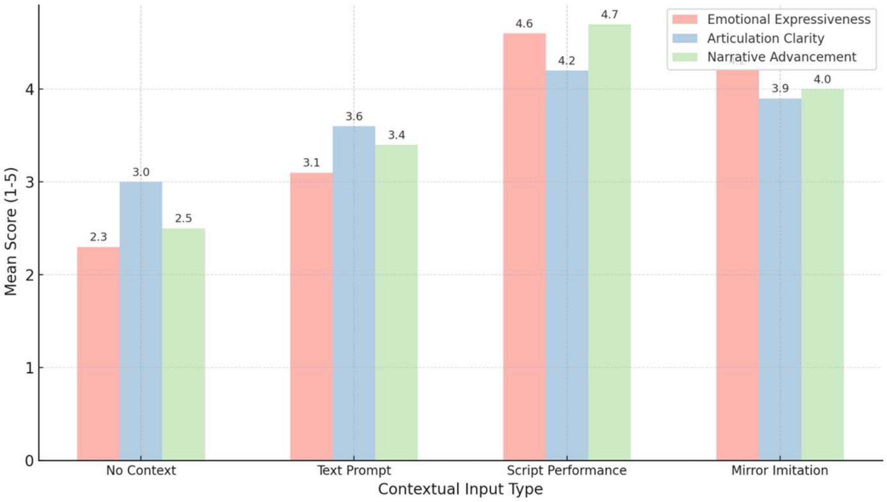


Figure 10: Effect of Contextual Guidance on Singing

Figure 11 The ablation study illustrates how different configurations of the feature enhancement module (FEM) and attention mechanism (SA) affect signal separation performance across three metrics: SDR, PESQ, and STOI. The complete model incorporating both FEM and SA (+FEM+SA) consistently outperforms all other setups, validating the synergy between multi-scale feature enhancement and attention-driven signal focus. This confirms that integrating both modules significantly improves separation quality in neural network-based audio processing.

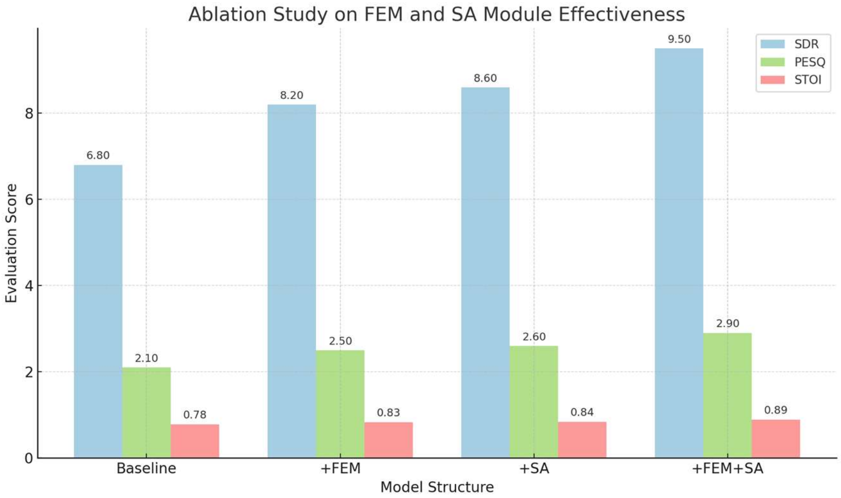


Figure 11: Ablation Study on FEM and SA Module Effect

The enhanced Figure 12 perceptual evaluation heatmap reveals clear trends in listener responses across original, vocal-only, and AI-reconstructed music versions, as evaluated by both musicians and general audiences. The original version consistently receives the highest scores across all dimensions—clarity, naturalness, and singing fidelity—serving as the perceptual benchmark. In contrast, the vocal-only version exhibits a marked decline in perceived quality, particularly in naturalness and fidelity, with musicians rating it more critically than the general audience. Notably, the AI-reconstructed version demonstrates substantial perceptual gains over the vocal-only output, especially in clarity and naturalness, with scores nearing those of the original, indicating that the integration of attention mechanisms (e.g., SA) and feature enhancement (FEM) significantly restores perceptual realism. The reconstructed output is particularly well-received by general listeners, suggesting its practical utility in mainstream audio applications. Overall, the results confirm that advanced deep learning-based reconstruction can approximate human-like perceptual quality and substantially improve user listening experience.

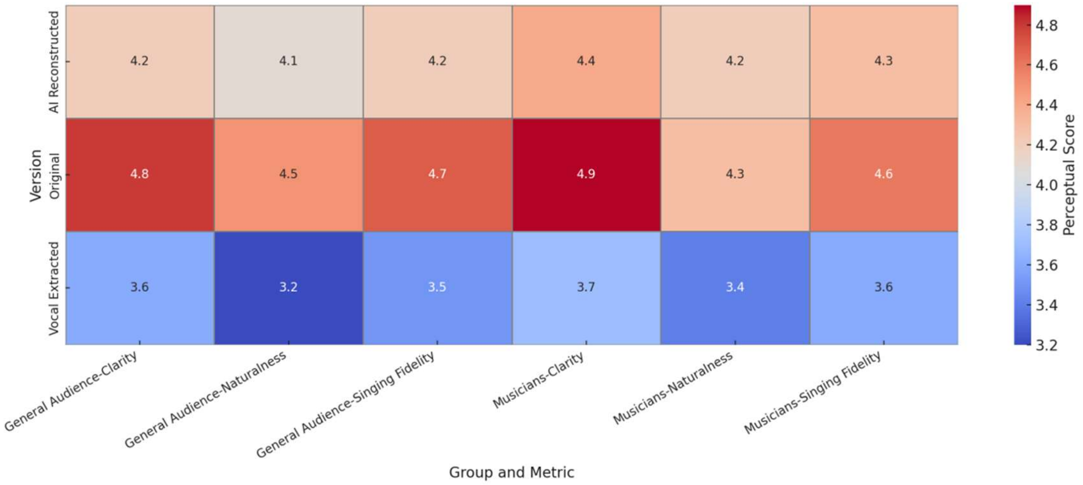


Figure 12: Perceptual Evaluation Heatmap

Figure 13 reveals clear advantages of constructivist and blended teaching methods over traditional approaches in musical theatre education. Across four key dimensions—expressiveness, role understanding, stage interaction, and improvisation—the constructivist group consistently outperforms others, especially in role immersion and improvisational fluency, reflecting the strengths of experiential, reflective learning. The blended method also demonstrates competitive results with relatively low variance, particularly excelling in stage interaction and improvisation, indicating its effectiveness in balancing structure and creativity. In contrast, the traditional group lags across all metrics, with pronounced underperformance in improvisational tasks, suggesting limitations in fostering adaptive and emotionally resonant performance. These findings underscore the pedagogical value of situational learning and support a shift toward constructivist-informed or hybrid instructional models in musical theatre training.

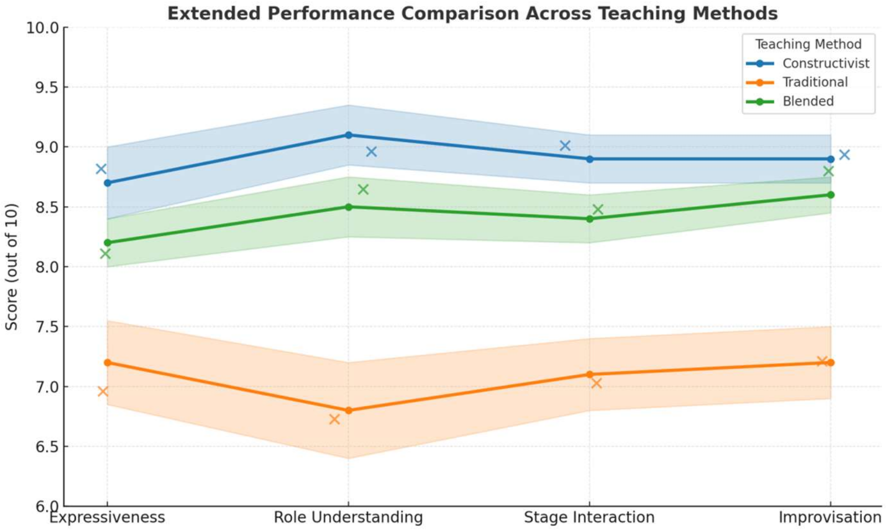


Figure 13: Extended Performance Comparison Across Teaching

IV. Conclusion

This study integrates constructivist pedagogy with advanced deep learning techniques to address both educational and technical challenges in musical theatre singing and music signal processing. Through a structured encoder-decoder model enhanced by self-attention and feature extraction modules, the proposed framework significantly improves the separation accuracy of vocal signals and the interpretability of musical expressions. Case studies confirm enhanced sub-module construction precision and more authentic signal reconstruction in both amplitude and phase domains. These findings not only support the use of intelligent neural models in audio processing but also demonstrate the potential for algorithm-assisted music education and artistic practice. Future work will extend the approach to real-time feedback and adaptive learning environments for vocal performance.

References

[1] Nassar, N., Jafar, A., & Rahhal, Y. (2020). A novel deep multi-criteria collaborative filtering model for recommendation system. Knowledge-Based Systems, 187, 104811.

[2] Al Jawarneh, I. M., Bellavista, P., Corradi, A., Foschini, L., Montanari, R., Berrocal, J., & Murillo, J. M. (2020). A pre-filtering approach for incorporating contextual information into deep learning based recommender systems. IEEE Access, 8, 40485-40498.

[3] Unger, M., Tuzhilin, A., & Livne, A. (2020). Context-aware recommendations based on deep learning frameworks. ACM Transactions on Management Information Systems (TMIS), 11(2), 1-15.

[4] Logesh, R., Subramaniyaswamy, V., Malathi, D., Sivaramakrishnan, N., & Vijayakumar, V. (2020). Enhancing recommendation stability of collaborative filtering recommender system through bio-inspired clustering ensemble method. Neural Computing and Applications, 32(7), 2141-2164.

[5] Yu, M., Quan, T., Peng, Q., Yu, X., & Liu, L. (2022). A model-based collaborate filtering algorithm based on stacked AutoEncoder. Neural Computing and Applications, 34(4), 2503-2511.

[6] Liu, H., Wang, Y., Peng, Q., Wu, F., Gan, L., Pan, L., & Jiao, P. (2020). Hybrid neural recommendation with joint deep representation learning of ratings and reviews. Neurocomputing, 374, 77-85.

[7] Chen, C., Zhang, M., Zhang, Y., Liu, Y., & Ma, S. (2020). Efficient neural matrix factorization without sampling for recommendation. ACM Transactions on Information Systems (TOIS), 38(2), 1-28.

[8] Ferrari Dacrema, M., Boglio, S., Cremonesi, P., & Jannach, D. (2021). A troubling analysis of reproducibility and progress in recommender systems research. ACM Transactions on Information Systems (TOIS), 39(2), 1-49.

[9] Jingchun Zhou, Jiaming Sun, Weishi Zhang, Zifan Lin. Multi-view underwater image enhancement method via embedded fusion mechanism. Engineering Applications of Artificial Intelligence, 2023, 121, 105946.

[10] Jingchun Zhou, Lei Pang, Weishi Zhang. Underwater image enhancement method by multi-interval histogram equalization. IEEE Journal of Oceanic Engineering, 48(2),2023: 474-488.

[11] Li, C., Kou, Y., Shen, D., Nie, T., & Li, D. (2024). Cross-grained Neural Collaborative Filtering for Recommendation. IEEE Access.

[12] Zhiwu, Wu . et al "Enhanced on-site testing of DC current transformers using improved EMD filtering and high-precision synchronization." Mari Papel y Corrugado Volume 2025: 8-15, doi:10.71442/mari2025-0002.