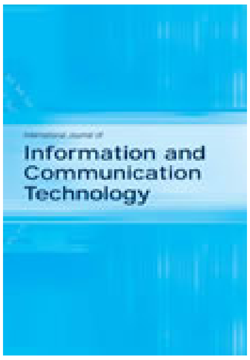
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of
Information and
Communication
Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Intelligent assessment system for singing skills based on time-frequency feature decoupling

Ying Zhang, Ruixue Sun, Hongrun Shao, Chunmeng Zhao

DOI: 10.1504/IJICT.2025.10075272

Article History:

Received:	04 August 2025
Last revised:	05 November 2025
Accepted:	05 November 2025
Published online:	12 January 2026

Intelligent assessment system for singing skills based on time-frequency feature decoupling

Ying Zhang, Ruixue Sun*, Hongrun Shao and Chunmeng Zhao

Arts Department,
Qinhuangdao Vocational and Technical College,
Qinhuangdao, 066100, China
Email: zhangying@qvc.edu.cn
Email: srx@qvc.edu.cn
Email: shaohongrun@qvc.edu.cn
Email: 13933528444@163.com
*Corresponding author

Abstract: Singing technique assessment is a crucial component in enhancing the quality of music education. To address the issue of insufficient assessment accuracy caused by the coupling of time-frequency features in existing methods, this paper first performs pre-processing on singing audio to extract time-frequency features. Then, by combining deep separable convolutions with dilated convolutions, it simultaneously models frequency and temporal features. Additionally, a residual network is employed to mitigate the gradient vanishing problem in deep network structures. Second, a spatio-temporal enhancement branch is constructed based on a bidirectional long short-term memory (BiLSTM) network. Through a gating mechanism, decoupled features are bidirectionally transmitted between temporal and frequency domains. Decoupled time-frequency feature sequences are then clustered to enable the model to intelligently evaluate singing segments. Experimental results show that the proposed model achieves at least a 4.71% improvement in evaluation accuracy, demonstrating a significant advantage over baseline models.

Keywords: singing technique evaluation; time-frequency feature decoupling; deep separable convolution; bidirectional long short-term memory model; feature clustering.

Reference to this paper should be made as follows: Zhang, Y., Sun, R., Shao, H. and Zhao, C. (2025) 'Intelligent assessment system for singing skills based on time-frequency feature decoupling', *Int. J. Information and Communication Technology*, Vol. 26, No. 51, pp.18–33.

Biographical notes: Ying Zhang received her Master's degree from the Hebei University in 2014. She is currently the Director and an Associate Professor at the Qinhuangdao Vocational and Technical College. Her research interests include music and dance.

Ruixue Sun received her Master's degree from Hebei University in 2016. She is currently a Lecturer at the Qinhuangdao Vocational and Technical College. Her research interests include music and dance.

Hongrun Shao received her Bachelor's degree from the Hebei Normal University in 2003. She is currently an Associate Professor of Qinhuangdao Vocational and Technical College. Her research interests include music education.

Chunmeng Zhao received her Bachelor's degree from Sichuan Conservatory of Music in 2002. She is currently a Lecturer at the Qinhuangdao Vocational and Technical College. Her research interests include musicology.

1 Introduction

In the field of music and art, singing carries rich emotions and cultural connotations. The degree of mastery of singing techniques directly affects the quality and artistic expressiveness of a performance (Zhang et al., 2024). Traditional singing technique evaluation primarily relies on the experience of vocal teachers, using auditory perception and visual observation to analyse technical elements such as pitch, rhythm, timbre, and resonance of performers (Yang et al., 2022). However, this evaluation method has obvious subjectivity and limitations. Manual evaluation finds it difficult to conduct a comprehensive analysis of subtle changes and complex features during singing, and is unable to provide detailed and quantifiable evaluation criteria, limiting the scientific nature and effectiveness of vocal education (Sear, 2024). However, singing acoustic signals have complex time-frequency coupling characteristics. Features from different techniques intertwine in the time and frequency domains, making it difficult for existing methods to achieve fine-grained technique decoupling and objective evaluation (Gallo, 2019). Therefore, an accurate and objective evaluation of singing techniques has significant practical significance. It not only helps learners promptly understand their strengths and weaknesses and adjust their training direction, but also provides quantitative basis for vocal teachers to optimise teaching methods and improve teaching efficiency.

Frič and Pavlechová (2020) extracted objective audio features representing timbre based on subjective perception scores, calculated their differences to obtain a timbre similarity matrix, and achieved the detection and evaluation of singing techniques. Ekici (2022) divided the music signals based on musical measures, inferred the overall features from partial features of the music signal, validated the chaotic characteristics in the music signal using the Lyapunov exponent, and evaluated the sound quality based on the detected music signal features. Chan (2018) used source characteristics, signal characteristics, sound field characteristics, auditory characteristics, and stereo sensation as the criteria layer, and ten sound quality evaluation elements as the plan layer, thereby constructing the singing technique evaluation index system structure. However, the accuracy of the assessment is not high.

Machine learning can extract acoustic features that are difficult for the human ear to capture (such as microsecond-level pitch fluctuations and harmonic distortions), generate evaluation results in real-time through the extraction of audio signal features, and significantly improve evaluation efficiency. Xu et al. (2022) extracted the MECC features of singing audio and used support vector machines (SVM) to output evaluation prediction results. Xia and Yan (2021) used prosody and singing quality features to classify singing techniques. At the first level, they first classified two different activation levels, and then used a Bayesian classifier for music teaching evaluation. Tang (2023) used multi-template Mel frequency cepstral coefficients (MFCC) clustering to label speech frames and employed a k-nearest neighbour algorithm based on prosody features to evaluate singing techniques, but the evaluation accuracy was not high. Traditional singing technique evaluation mainly relies on experience-based judgment, analysing technical elements such as pitch, rhythm, and timbre through auditory and visual observations, leading to low precision in evaluations.

The above research mostly relies on manually designed audio features, which often fail to fully capture the complex connotations of singing techniques, leading to large prediction errors. Although machine learning-based singing skill assessment models can extract audio signal features in real-time, current feature extraction methods may fail to comprehensively capture all key elements of singing ability. Furthermore, machine learning models can only make objective judgments based on data features and are unable to capture these abstract attributes. Deep learning methods integrate deep neural networks into the feature extraction system, perform deep learning feature extraction on spectrograms, learn significant representations of singing techniques, and outperform machine learning algorithms. Donati et al. (2023) used a pre-trained AlexNet model to learn deep feature representations from three-channel speech Mel spectrograms and adopted a linear SVM for classification, improving evaluation accuracy. Liu and Hui (2022) designed a multi-scale convolutional neural network (CNN) to extract multimodal features of singing techniques and used an attention mechanism for feature fusion, achieving an evaluation accuracy of 79.45%. Pati et al. (2018) modelled the audio map of singing techniques as a spatiotemporal graph structure and designed a temporal graph convolutional network (GCN) for feature extraction, but such methods rely on static adjacency matrices that struggle to capture the dynamic associations of audio features. Li and Zhang (2022) combined transfer learning with multi-head self-attention from the Transformer to jointly model the spatial relationships and temporal evolution patterns of time-frequency features of singing techniques, but their computational complexity increases significantly with sequence length. To resolve the above contradiction, feature decoupling has become a new technological breakthrough. Guo and Tang (2023) used i-vector (Ibrahim and Ramli, 2018) to represent the timbre features of singers and voice posteriori graph to represent singer-independent singing features, to achieve feature decoupling, and used the decoupled features as inputs of the fully connected layer to obtain evaluation results. Chang (2025) employed two decoupled masked autoencoders to separately extract time-domain and frequency-domain features from singing audio and designed a multi-level feature fusion strategy, significantly improving evaluation accuracy.

According to an in-depth analysis of current research, existing studies still face multiple challenges. Most methods can extract time-frequency features, but lack an effective multi-scale modelling mechanism, especially when processing different time scales, they fail to effectively integrate short-term and long-term information, and the spatiotemporal decoupling weakens feature interaction. To address the above issues, this paper proposes an intelligent singing technique evaluation model based on time-frequency feature decoupling. First, singing audio pre-processing is completed by utilising audio sampling, normalisation, framing, and time-frequency transformation processes to extract the time-frequency features of the audio. Variational autoencoders (VAE) are used to represent the time-frequency features of singing audio. Then, a multi-scale time-frequency feature enhancement module is designed. It combines depthwise separable convolution and dilated convolution, enabling simultaneous modelling of frequency domain features and long-term periodic behaviour patterns. By embedding a mixed attention mechanism, the semantic responses between time-frequency feature channels are further enhanced. At the same time, residual networks with attention mechanisms are used to alleviate the gradient attenuation problem in deep network structures. Furthermore, time-frequency feature decoupling is achieved based on bidirectional gates, and a spatiotemporal enhancement branch is constructed using a BiLSTM network. Through a gating mechanism, decoupled features are bidirectionally transmitted in time and frequency domains. The decoupled time-frequency feature sequences are clustered to enable the model to intelligently evaluate different singing segments. Experimental outcome indicates that the evaluation accuracy of the proposed model is 93.48%, which is at least 4.71% higher than that of baseline models, significantly improving the evaluation accuracy and providing new insights for the automation and precision of singing technique evaluation.

2 Relevant technologies

2.1 Feature decoupling representation learning theory

Due to the lack of explicit signal supervision in the deep learning process, different features are likely to be coupled, which makes it difficult for the model to distinguish the important information required for specific tasks (Zhou et al., 2020). Unlike controlling shallow, single-dimensional features with relatively clear semantic information, in singing technique quality evaluation tasks, a more complex high-dimensional feature space needs to be decomposed so that different feature components can independently learn semantic information corresponding to different subtasks or training objectives, thereby improving the model's generalisation ability, interpretability, and task adaptability. Conventional decoupling methods like EMD primarily target signal processing domains, demonstrating strong performance on specific data types such as one-dimensional time series signals. However, they face limitations when handling multimodal data like images, text, and speech. Feature decoupling representation learning, in contrast, exhibits broader adaptability. For instance, in the image domain, it can decompose images into independent features like shape, colour, and texture. In speech processing, it can separate features such as pitch, timbre, and intonation. In text analysis, it can decompose features like semantics, syntax, and sentiment. This cross-domain adaptability makes feature-based decoupling representation learning highly promising for applications across multiple fields.

Feature decoupling representation learning aims to learn a representation from data such that different dimensions of this representation correspond to independent varying factors in the data generation process. Feature decoupling representation learning aims to find a specific latent feature representation in which each latent feature captures data variations independent of other latent features, i.e., each latent feature controls an independent factor of the data (Wang et al., 2024a). Most feature decoupling representation learning methods are based on the VAE architecture (Sewak et al., 2020), optimised by maximising the evidence lower bound, with the loss function expressed as follows.

$$L_{VAE} = E_{q_\phi(h|x)}\left[\ln p_\theta(x \mid h)\right] - D_{KL}\left[q_\phi(h \mid x) \| p(h)\right] \tag{1}$$

where x is the input data and h is the hidden layer feature. The first term in L_{VAE} is the reconstruction loss, and the second term aims to approximate the prior distribution $p(h)$ of the latent variables using the parameterised posterior distribution $q_\phi(h|x)$. The generative model $p_\theta(x|h)$ is the VAE decoder parameterised by the neural network.

2.2 Attention mechanism

The attention mechanism is designed to address long-range dependencies in sequence data and improve the network's memory. In traditional sequence models, such as recurrent neural networks (RNN), long-range dependencies often cause problems like vanishing or exploding gradients, making it difficult for the model to effectively capture distant information. To solve this issue, the attention mechanism was introduced, and by incorporating learnable weights, the model can dynamically adjust the degree of attention to different parts based on the relevance of different positions in the input sequence (Yu et al., 2020). This mechanism enables the model to capture long-range dependencies between inputs and outputs, which is crucial for many natural language processing tasks (Song et al., 2020). For example, in speech generation, the acoustic information required for the model to generate each phoneme often only focuses on a short context window around that phoneme.

Given a query, different keys are selected from the source and their relevance to the query is computed; different values corresponding to these keys receive different weight coefficients. Finally, the attention value is calculated through the weighted average of the results. The calculation formula is as follows, where $Lx = ||Source||$ is the length of the source, and $Value_i$ is the value.

$$Attention(Query, Source) = \sum_{i=1}^{Lx} Similarity(Query, Key) * Value_i \tag{2}$$

According to the different domains of the attention mechanism, it can be categorised into types such as channel attention (CAM), spatial attention (SAM), and hybrid attention (CBAM) (Wang et al., 2024b). CAM improves the model's focus on input features, with a simple concept and fewer parameters, making it easily extensible to other networks. SAM focuses on the spatial location information in feature maps, aiming to learn the importance of different spatial positions in the feature map so that the model can focus on important regions in the image while ignoring irrelevant areas. CBAM integrates attention models from multiple dimensions. The most common approach is combining CAM and SAM, enabling the network to adaptively focus on important regions in the entire feature map.

3 Design of an adaptive selection mechanism for multi-modal sensor data based on mutual information

3.1 Singing audio data pre-processing

To improve the evaluation accuracy of the singing skill assessment system, it is essential to ensure the quality of the input variables. This paper completes the pre-processing of singing audio through audio sampling, normalisation, framing, and time-frequency transformation to determine the time-frequency features of the singing audio. Based on this, the time-frequency features of the singing audio are represented using a VAE, laying the foundation for the subsequent establishment of the evaluation model.

The singing audio pre-processing process includes audio sampling, normalisation, framing, and time-frequency transformation. Singing audio signals exhibit short-term stationary characteristics, so a Hamming window is selected for the framing and windowing processing of the singing audio signals, setting the number of sampling points per frame to *n*. The time-frequency transformation of the singing audio signal uses the short-time Fourier transform.

Singing audio consists of various notes with certain time durations, where each note's main characteristic is a relatively stable spectrum. This indicates that the notes within the singing audio appear as a series of spectral segments on a spectrogram, featuring significant differences between segments and minimal differences within segments. Based on this, the distance measurement algorithm can be selected to perform note segmentation processing. As a distance measurement method that integrates the mean and method of data segments, the distance measurement algorithm can determine the differences between singing audio segments. Setting the data window length to five frames, the distance measurement algorithm DIS is described by equation (3), as shown below.

$$DIS = \frac{(\mu_1 + \mu_2)^T (\mu_1 + \mu_2)^T (\mu_1 + \mu_2)}{tr(\Sigma_1) + tr(\Sigma_2)} \tag{3}$$

where μ_1 is the mean vector of the previous singing audio segment's features, μ_2 is the mean vector of the next singing audio segment's features, $tr(\Sigma_1)$ is the trace of the covariance matrix of the previous singing audio segment's features, and $tr(\Sigma_2)$ is the trace of the covariance matrix of the next singing audio segment's features. Under conditions where there is a significant difference in feature means between two singing audio segments and minimal variance in feature means within segments, the distance measurement can describe the distance between the two singing audio segments, and the two are directly proportional.

The short-time amplitude spectrum is used to determine feature parameters. Through slid windows sliding data window according to frames, the distance measurement function $DIS(t)$ for frame number t is determined.

$$DIS(t) = \frac{(\mu_{t,1} + \mu_{t,2})^T (\mu_{t,1} + \mu_{t,2})}{tr(\Sigma_{t,1}) + tr(\Sigma_{t,2})} \tag{4}$$

where $\mu_{t,1}$ is the mean vector of the features of the two singing audio segments before frame t, $\mu_{t,2}$ is the mean vector of the features of the two singing audio segments after frame t, $tr(\Sigma_{t,1})$ is the trace of the covariance matrix of the features of the two singing audio segments before frame t, and $tr(\Sigma_{t,2})$ is the trace of the covariance matrix of the features of the two singing audio segments after frame t.

Calculate all maximum points within $DIS(t)$, set the threshold for the mean of $DIS(t)$ to T_1, and remove the maximum points in which value $<T_1$. At the same time, in fast-paced singing, the duration of a quarter note is roughly 1/2 s. Considering that the duration of an eighth note and a sixteenth note are 1/2 and 1/4 of the quarter note, respectively, the segment distance must be set to more than 100 ms. Otherwise, the

corresponding maximum points must be eliminated, leaving the remaining maximum points as the note cut points. Since singing audio includes both voiced segments and non-voiced segments, the voiced segments and non-voiced segments must be judged using a voiced segment detection algorithm after cutting. The spectral variance of voiced segments is significantly greater than that of non-voiced segments. Therefore, the spectral variance can be used as a feature parameter to identify the voiced segments in singing audio.

3.2 Time-frequency feature representation of singing audio data

The above singing audio pre-processing process can effectively reduce the false alarm rate of melody localisation in singing audio. Note splitting is achieved using a distance measurement algorithm (Casey et al., 2008), and the voiced segments in electronic music are determined using the variance method. On this basis, the Viterbi algorithm (Rao et al., 2016) is used to track the dominant fundamental frequency trajectory of the voiced segments, while a fundamental frequency discrimination model is used to determine the main melody of electronic music, thereby obtaining the time-frequency data of the singing audio. In order to facilitate subsequent model recognition and evaluation, the time-frequency data of the singing audio needs to be feature represented. This paper uses a probabilistic inference model based on VAE, aiming to learn the time-frequency features of singing audio data and represent them.

The time-frequency feature encoding consists of a time-frequency inference network and a time-frequency generation network. The time-frequency inference network models the Mel-frequency cepstral coefficients *U* by adjusting the input structure of the CNN. It extracts the time-frequency features of *U* and encodes them into a time-frequency latent representation Z_f^1. The spectral generation network takes Z_f^1 as input and generates a reconstructed feature vector U'_f.

$$\mu_U, \ln \sigma_U^2 = E_f(U) \tag{5}$$

$$Z_f^1 = \mu_U + \sigma_U \odot \varepsilon \tag{6}$$

$$U'_f = D_f\left(Z_f^1\right) \tag{7}$$

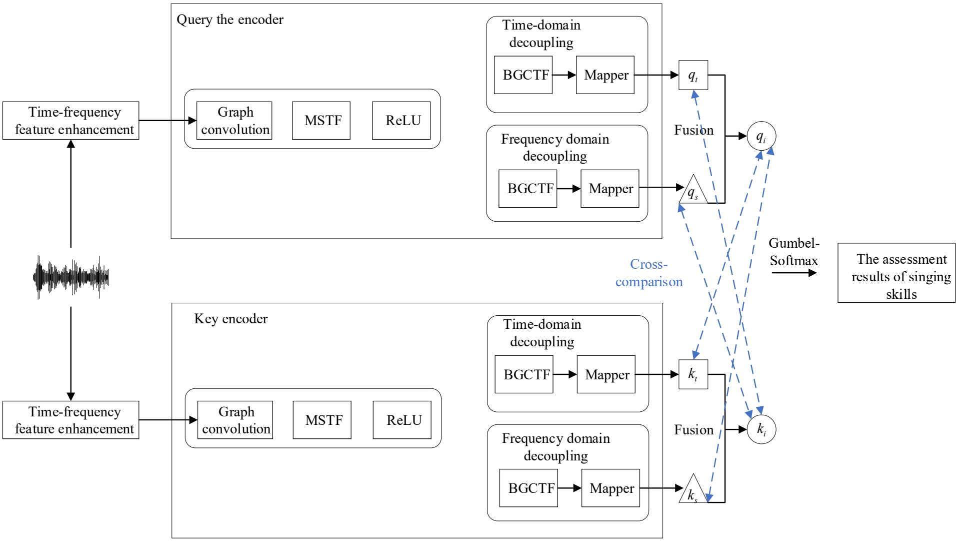
where $E_f(*)$ is the time-frequency inference network, $\varepsilon \sim \mathcal{N}(0,1)$ is a random noise sampled during the reparameterisation process, ⊙ is element-wise multiplication, $Z \sim N(\mu_U, \sigma_U^2),$, and $D_f(*)$ is the spectral generation network, where U'_f is the reconstructed vector.

4 Intelligent evaluation of singing techniques by decoupling time-frequency features

4.1 Multi-scale time-frequency feature enhancement

To address the insufficiency of time-frequency feature modelling and multi-scale fusion in singing skill evaluation, this paper proposes an intelligent singing skill evaluation system based on time-frequency feature decoupling. The system model is mainly composed of enhanced multi-scale time-frequency features (MSTF), time-frequency feature decoupling module (BGCTF), and singing skill element score prediction module, as indicated in Figure 1. First, it combines dilated convolution with deep separable temporal convolution to establish a cross-scale feature dynamic interaction mechanism, solving the scale mismatch problem caused by fixed dilation rates in traditional methods. Then, it embeds the CBAM mechanism and designs a residual network with attention mechanism to strengthen the semantic response between singing skills and feature channels while alleviating the gradient vanishing problem. Next, a time-frequency feature decoupling enhancement branch is constructed based on a BiLSTM network, achieving bidirectional dynamic modelling of the vocal spectrogram and time-frequency features through a gating mechanism. Finally, the Gumbel-Softmax clustering is used to predict and score the singing skills, obtaining the final intelligent singing skill evaluation results.

Figure 1 Intelligent assessment model for singing techniques based on the decoupling of time-frequency features (see online version for colours)



For the goal of coping with the issues of insufficient capture of complex time-frequency patterns and inefficient feature fusion in singing skill evaluation, this paper proposes a MSTF module based on an attention mechanism. The module achieves efficient modelling of long-term and short-term time-frequency feature dependencies through a hierarchical feature enhancement and dynamic fusion mechanism. First, traditional studies use standard 3D convolutions for feature extraction, which suffer from limited receptive fields. To address this, this paper proposes combining deep separable convolutions with dilated convolutions. First, the standard convolution is decomposed into a cascade operation of depthwise convolution and pointwise convolution, and a divide-and-conquer strategy is used for feature extraction, as shown in equations (8) and (9).

$$F_{mid} = \text{ReLU}\left(BN\left(P_{1\times 1}\left(F_{in}\right)\right)\right) \tag{8}$$

$$F_{out} = P_{1\times 1}\left(DWConv\left(F_{mid}\right)\right) \tag{9}$$

where $P_{1\times 1}$ is the pointwise convolution of 1 × 1, F_{in}, F_{mid}, and F_{out} are the input features, intermediate features, and output features, respectively. BN is the batch normalisation operation, ReLU is the activation function, and DWConv is the depthwise convolution. Next, a dilation rate sequence is introduced in the temporal dimension to construct multi-level temporal receptive fields. The model can capture context information of different granularities simultaneously through different dilation rates. A small dilation rate focuses on fine-grained features between adjacent frames for capturing local features, whereas a large dilation rate spans multiple frame time ranges to capture action periodic characteristics. The actual coverage range RF of depthwise separable convolution with kernel size k and dilation rate d is calculated as follows.

$$RF = k + (d-1)\times(k-1) \tag{10}$$

Next, to optimise the multi-branch feature fusion process, the CBAM mechanism is introduced, as shown in Figure 2. This component includes parallel max-pooling and average-pooling paths. Channel weight distribution is learned through a fully linked level, and the equation is as follows.

$$w_c = \sigma\left(MLP\left(Avg(F)\right) \oplus MLP\left(\max(F)\right)\right) \tag{11}$$

where F is the input feature, σ is the Sigmoid function, ⊕ is element-wise addition, MLP represents the fully connected layer, Avg and Max respectively represent the average pooling layer and the max pooling layer, and w_c is the channel weight. After the channel weight calculation, the size of the attention weight map is dynamically adjusted to adapt to different spatial dimensions of the input, solving the feature map size mismatch problems caused by the pooling operation. Subsequently, the channel weighting is applied to the original feature map through the element-wise product ⊗.

Finally, to address the gradient attenuation problem in deep networks, a residual structure incorporating an attention mechanism is constructed.

$$F_{res} = AttnConv\left(P_{stride}\left(F_{in}\right)\right) \tag{12}$$

where $AttnConv$ is the temporal convolution layer embedded with CBAM, P_{stride} is the stride-1 × 1 convolution, F_{in} and F_{res} respectively represent the input and residual features. The input feature is first compressed along the channel dimension through the stride 1 × 1 convolution to reduce the computational complexity. Subsequently, the temporal convolution layer of CBAM dynamically enhances key features related to singing skill assessment. Finally, the processed features are added to the high-frequency features output from the main path to achieve stable gradient propagation.

4.2 Time-frequency feature decoupling based on bidirectional gating

To overcome the problem of insufficient one-way information flow and feature interaction after traditional time-frequency feature decoupling, this article proposes a BGCTF module, which realises the decoupling, fusion, and enhancement of time-frequency features through a hierarchical bidirectional interaction mechanism. The core innovation lies in using BiLSTM (Wijaya et al., 2024) to build a dynamic interaction mechanism for time-frequency features.

First, the extracted time-frequency features u_t are rearranged into v_t through a specific dimensional reconstruction strategy. Second, a BiLSTM structure is proposed to build a bidirectional feature propagation mechanism. The temporal encoder captures the forward and backward dependencies of the audio through bidirectional gating units, modelling the dynamic patterns of singing audio, where the forward LSTM captures historical trends and the backward LSTM infers future singing intentions. The concatenation of their hidden states forms a dynamically enhanced temporal context. In the frequency domain decoupling branch, the frequency domain encoder utilises a memory gating mechanism to learn the implicit topological constraint relationships between joints. Forward propagation encodes the local joint collaboration patterns, while backward propagation captures the time-frequency feature correlations of singing audio, outputting a spatial topological context.

Finally, max pooling is applied to the output of BiLSTM to compress the dimensionality. v_t extracts salient features along the time axis, and v_s captures key frequency domain features along the frequency axis. This strategy retains the most discriminative features of each modality. Ultimately, the bidirectional characteristics of v_t and v_s completely preserve the contextual information of the sequence.

Through the bidirectional gating mechanism of BiLSTM, the sequential dependency differences of traditional methods are overcome. The gating units can adaptively adjust the interaction strength of time-frequency features, enhancing the model's robustness to noise while reducing the risk of overfitting.

4.3 Score prediction for singing technique elements

After obtaining the decoupled time-frequency features, this paper clusters the time-frequency feature sequences of singing audio to enable the model to perform intelligent evaluation of different singing segments. In the absence of sequential and category labels, an unsupervised clustering-based segmenting method is used, specifically performing differentiable Gumbel-softmax clustering (Chaudhary and Singh, 2023) on the feature of each segment.

Specifically, let there be K randomly initialised cluster centres. In practice, referring to the evaluation rules of singing techniques, singing techniques are divided into three categories: basic vocalisation, pitch and rhythm control, and timbre and expressiveness, so K is directly set to 3, denoted as $C_1, \ldots, C_K$. The distance from each element f_t in the decoupled time-frequency features F_t to each centre is then computed accordingly.

$$d_i = \left\| f_T - C_i \right\|^2 \tag{13}$$

Subsequently, a random vector $g_i \sim Gumbel(0, 1)$ is added to simulate the randomness of discrete random sampling, and a hyperparameter $\tau_{cluster}$ is used to control the smoothness of the clustering, as shown below, where the random variable u_i is sampled from a uniform distribution $U(0, 1)$.

$$P_i = \text{softmax}\left(\frac{d_i + g_i}{\tau_{cluster}} \right) \tag{14}$$

$$g_i = -\log\left(-\log\left(u_i\right)\right) \tag{15}$$

After that, a category vector is obtained by weighting according to the probability distribution, as shown below.

$$z = \sum_{i=1}^{K} P_i \cdot C_i \tag{16}$$

$\tau_{cluster}$ usually takes small values to make z closer to the cluster centres rather than the original features, enhancing its ability to represent categories and reducing direct interference from the semantics of the features. Finally, a linear transformation is applied to the clustering results to obtain z', which adds extra information to each feature to indicate its category, and then updates the feature representation with category information through residual connections (i.e., $f_T' \leftarrow f_T + z'$) to assist subsequent models in recognising their semantics. The sequence F_T' with added category information is divided into several segments, which are then sequentially fed into an MLP for computation. Ultimately, the results of all singing segments are summed to obtain the predicted score for the intelligent assessment $\hat{y}_T$.

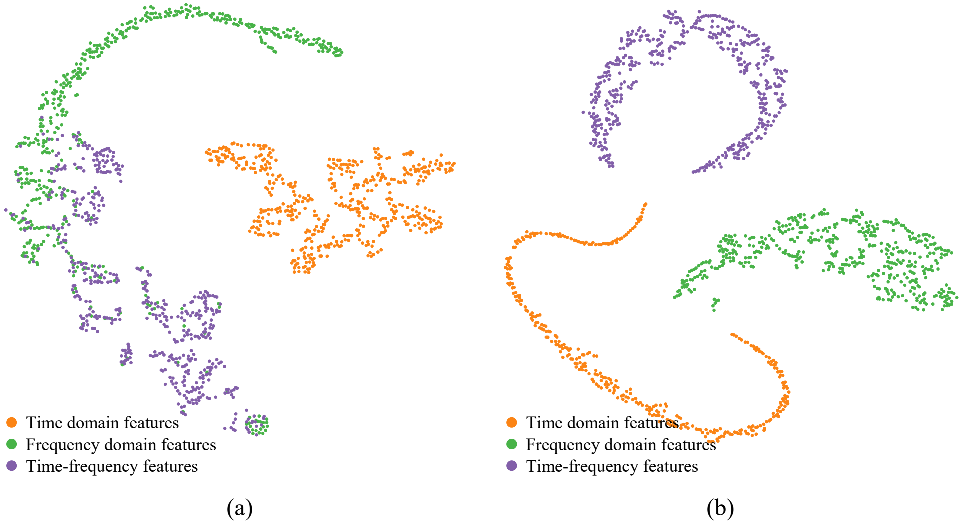
5 Experimental results and analyses

5.1 Analysis of intelligent assessment results for singing techniques

This paper uses the Tianqin Singing Evaluation Dataset released by Tencent Music's Tianqin Lab in collaboration with Tsinghua University and other institutions. This dataset includes 1,000 dry vocal tracks of 10 songs, with accompanying Musical Instrument Digital Interface (MIDI) and lyric files, suitable for singing technique evaluation research. In this dataset, 70% of the data was randomly selected as the training set, and 30% was selected as the test set. The experiments are implemented using the PyTorch 1.13.0 framework with an RTX 4090 GPU. This paper sets the dataset batch size batch_size to 64 and the initial hyperparameter for softmax to 0.2. During training, stochastic gradient descent with momentum 0.9 and weight decay 0.0001 was used for optimisation. In order to fully learn the features in the dataset and ensure the learning rate stabilises near the optimal, the training epoch is set to 450.

This paper uses t-SNE to visualise the differences between the time-frequency features before and after decoupling, as shown in Figure 2. Similar features show no significant distribution differences, so the t-SNE does not distinguish them clearly and therefore they are not displayed. From Figure 2, it is evident that the untreated difference features exhibit significant distinctions in time-domain features, frequency-domain features, and time-frequency features, which can be attributed to the inherent differences between the different singing audio segments. Additionally, some overlap exists between the frequency-domain features and time-frequency features, indicating that consistent information exists in the difference features. The difference features after decoupling are not only successfully separated but also appear more compact, indicating that TFFD effectively extracts unique information between time-frequency features.

Figure 2 Visualisation of time-frequency characteristics before and after decoupling, (a) before decoupling (b) after decoupling (see online version for colours)



5.2 Comparative experiments

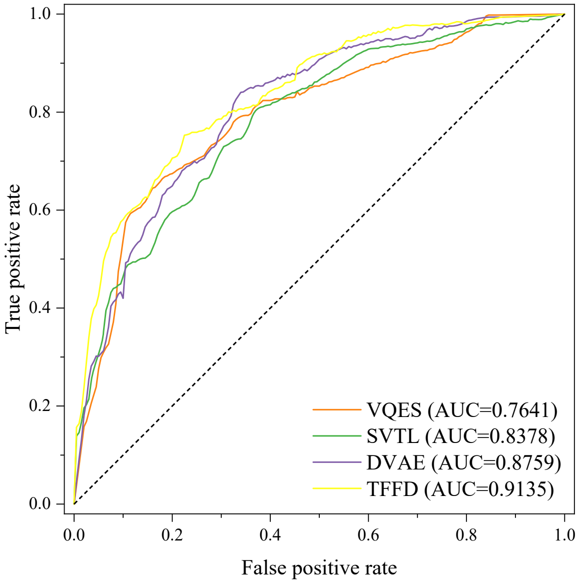
The experiment selects VQES (Liu and Hui, 2022), SVTL (Li and Zhang, 2022), and DVAE (Chang, 2025) as baseline models. The proposed model is denoted as TFFD. The evaluation accuracy of the basic voice production, pitch and rhythm control, and tone and expressiveness singing skills for different models is shown in Table 1. TFFD achieves an evaluation accuracy of 94.62% for the basic voice production singing skills, which is an improvement of 13.14%, 10.71%, and 5.56% compared to VQES, SVTL, and DVAE, respectively. By comparing the singing skills of pitch and rhythm control, TFFD achieves an evaluation accuracy improvement of 14.02%, 6.49%, and 4.36% compared to QES, SVTL, and DVAE, respectively. The average evaluation accuracy of VQES, SVTL, DVAE, and TFFD for these three singing skills is 79.45%, 83.9%, 88.77%, and 93.48%, respectively. The average evaluation accuracy of TFFD is at least 4.71% higher than the other three models. Numerical analysis of the above experimental results indicates that TFFD can achieve accurate assessment of various singing skills.

Table 1 The assessment accuracy rates of different types of singing techniques

Model	*VQES*	*SVTL*	*DVAE*	*TFFD*
Basic voice production	81.48%	83.91%	89.06	94.62%
Pitch and rhythm control	77.51%	85.04%	87.17	91.53%
Tone and expressiveness	79.36%	82.75%	90.08	94.29%

The ROC curves of different models are compared as shown in Figure 3. The AUC values of VQES, SVTL, DVAE, and TFFD are 0.7641, 0.8378, 0.8759, and 0.9135, respectively. Compared to the baseline model, the AUC value of TFFD improved by 4.29-19.55%. VQES designs multi-scale CNNs and CNNs to evaluate and predict vocal techniques, but the scale design relies on prior knowledge, such as low scales capturing pitch details and high scales capturing rhythmic patterns. However, vocal technique evaluation needs to cover multiple dimensions such as pitch accuracy, rhythm, and timbre, and fixed scales are difficult to dynamically adapt. Although SVTL introduced a Transformer with parallel computing for intelligent evaluation of vocal techniques, vocal technique evaluation needs to simultaneously focus on absolute and relative sequences. The attention weights of the Transformer only reflect absolute relationships between positions and are difficult to model temporal dynamics. Although DVAE considers the decoupling of time-frequency features in audio, the evaluation of vocal techniques must consider temporal dynamics, such as the frequency changes of vibratos and the intensity control of crescendos and decrescendos. A standard autoencoder lacks explicit modelling of temporal relationships. In summary, TFFD models the time domain and frequency domain enhancement branch based on BiLSTM, and passes the features after time-frequency decoupling bidirectionally through the time domain and frequency domain branches via a gating mechanism, thereby effectively filtering noise and enhancing context dependency. Through these designs, TFFD can significantly improve the accuracy of evaluating vocal techniques.

Figure 3 The ROC curves of different models (see online version for colours)



5.3 *Ablation experiment*

This section verifies the effectiveness of the MSTF module and the BGCTF model through ablation experiments. The experimental results are shown in Table 2. In the table, TFFD/MSTF indicates removing the MSTF module while retaining the BGCTF module. TFFD/BGCTF indicates removing the BGCTF module while retaining the MSTF module. TFFD/(MSTF + BGCTF) indicates removing both the MSTF and BGCTF modules, and directly using the pre-processed time-frequency feature representation as input to the evaluation model.

Table 2 The ablation experiment results of the TFFD model

Model	*TFFD/MSTF*	*TFFD/BGCTF*	*TFFD/(MSTF + BGCTF)*	*TFFD*
Accuracy	80.33%	75.61%	72.51%	93.48%
AUC	0.8205	0.7843	0.7469	0.9135

TFFD/(MSTF+BGCTF) shows the lowest evaluation accuracy and AUC value, indicating that MSTF and BGCTF have the greatest impact on the performance of the evaluation model; only through their synergistic effect can optimal performance be achieved. The evaluation accuracy and AUC of TFFD are 93.48% and 0.9135, respectively, which are improved by 13.15% and 11.33% compared to TFFD/MSTF, and by 17.87% and 16.47% compared to TFFD/BGCTF. The results show that MSTF effectively enhances the representational ability of singing details and the overall structure, while BGCTF significantly improves the time-frequency feature correlation modelling of singing audio. This fully verifies the importance of multi-scale time-frequency features and long-term dependency modelling, and further highlights the generalisation and robustness of the module architecture.

6 Conclusions

Intelligent evaluation of singing skills plays a crucial role in music education, vocal research, and singing training. To address the issue where current research fails to effectively integrate short-term and long-term information, leading to weak feature interactions due to spatiotemporal decoupling, this paper proposes an intelligent singing skill evaluation model based on time-frequency feature decoupling. First, singing audio pre-processing is performed through audio sampling, normalisation, framing, and time-frequency transformations to extract the time-frequency features of the audio. A VAE is then used to represent the time-frequency features of the singing audio. Subsequently, by introducing a multi-scale time-frequency feature enhancement module that combines depthwise separable temporal convolution networks with dilated convolution scale selection, short-term and long-term periodic audio features can be perceived simultaneously, significantly enhancing the capacity for multi-scale feature fusion. Furthermore, to further improve the modelling ability of time-frequency features, this paper also designs a time-frequency feature decoupling module based on bidirectional gating. This module uses BiLSTM for time-domain and frequency-domain enhancement branch modelling and employs a gating mechanism to bidirectionally transmit decoupled time-frequency features in both time and frequency domain branches, thereby effectively filtering noise and enhancing contextual dependencies. Finally, clustering is applied to the decoupled time-frequency feature sequences to enable the model to perform intelligent evaluation of different singing segments. Experimental results show that the evaluation accuracy of the proposed model is 93.48%, with an AUC of 0.9135, achieving precise intelligent assessment of singing skills.

Declarations

All authors declare that they have no conflicts of interest.

References

Casey, M., Rhodes, C. and Slaney, M. (2008) 'Analysis of minimum distances in high-dimensional musical spaces', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 5, pp.1015–1028.

Chan, H. (2018) 'A method of prosodic assessment: insights from a singing workshop', *Cogent Education*, Vol. 5, No. 1, pp.14–21.

Chang, D. (2025) 'Vocal performance evaluation of the intelligent note recognition method based on deep learning', *Scientific Reports*, Vol. 15, No. 1, pp.13–27.

Chaudhary, L. and Singh, B. (2023) 'Gumbel-SoftMax based graph convolution network approach for community detection', *International Journal of Information Technology*, Vol. 15, No. 6, pp.3063–3070.

Donati, E., Chousidis, C., Ribeiro, H.D.M. and Russo, N. (2023) 'Classification of speaking and singing voices using bioimpedance measurements and deep learning', *Journal of Voice*, Vol. 4, pp.37–45.

Ekici, T. (2022) 'An evaluation on the human voice and the act of singing', *Turkish Online Journal of Educational Technology*, Vol. 21, No. 3, pp.1–14.

Frič, M. and Pavlechová, A. (2020) 'Listening evaluation and classification of female singing voice categories', *Logopedics Phoniatrics Vocology*, Vol. 45, No. 3, pp.97–109.

Gallo, D.J. (2019) 'Formative assessment practices and children's singing accuracy: a mixed methods inquiry', *International Journal of Music Education*, Vol. 37, No. 4, pp.593–607.

Guo, Y. and Tang, Y. (2023) 'The assessment model of online vocal music teaching quality under the optimized DL model', *Intelligent Systems with Applications*, Vol. 20, pp.20–36.

Ibrahim, N.S. and Ramli, D.A. (2018) 'I-vector extraction for speaker recognition based on dimensionality reduction', *Procedia Computer Science*, Vol. 126, pp.1534–1540.

Li, R. and Zhang, M. (2022) 'Singing-voice timbre evaluations based on transfer learning', *Applied Sciences*, Vol. 12, No. 19, pp.31–42.

Liu, Y. and Hui, D. (2022) 'Voice quality evaluation of singing art based on 1DCNN model', *Mathematical Problems in Engineering*, Vol. 20, No. 1, pp.74–81.

Pati, K.A., Gururani, S. and Lerch, A. (2018) 'Assessment of student music performances using deep neural networks', *Applied Sciences*, Vol. 8, No. 4, pp.50–67.

Rao, Z., Guan, X. and Teng, J. (2016) 'Auto chord recognition based on sparse representation classification and Viterbi algorithm', *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 11, No. 11, pp.189–198.

Sear, J. (2024) 'Modern vocal pedagogy: investigating a potential curricular framework for training popular music singing teachers', *Journal of Popular Music Education*, Vol. 8, No. 2, pp.239–254.

Sewak, M., Sahay, S.K. and Rathore, H. (2020) 'An overview of deep learning architecture of deep neural networks and autoencoders', *Journal of Computational and Theoretical Nanoscience*, Vol. 17, No. 1, pp.182–188.

Song, G., Wang, Z., Han, F., Ding, S. and Gu, X. (2020) 'Music auto-tagging using scattering transform and convolutional neural network with self-attention', *Applied Soft Computing*, Vol. 96, pp.10–22.

Tang, H. (2023) 'Evaluation method of singing pronunciation quality based on artificial intelligence technology', *Procedia Computer Science*, Vol. 228, pp.526–532.

Wang, W., Li, Z., Liu, S., Zhang, L., Yang, J. and Wang, Y. (2024a) 'Feature decoupling and interaction network for defending against adversarial examples', *Image and Vision Computing*, Vol. 144, pp.32–45.

Wang, Y., Wang, W., Li, Y., Jia, Y., Xu, Y., Ling, Y. and Ma, J. (2024b) 'An attention mechanism module with spatial perception and channel information interaction', *Complex & Intelligent Systems*, Vol. 10, No. 4, pp.5427–5444.

Wijaya, N.N., Setiadi, D.R.I.M. and Muslikh, A.R. (2024) 'Music-genre classification using bidirectional long short-term memory and Mel-frequency cepstral coefficients', *Journal of Computing Theories and Applications*, Vol. 1, No. 3, pp.243–256.

Xia, X. and Yan, J. (2021) 'Construction of music teaching evaluation model based on weighted Naïve Bayes', *Scientific Programming*, Vol. 10, No. 2, pp.71–79.

Xu, Y., Wang, W., Cui, H., Xu, M. and Li, M. (2022) 'Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy', *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 20, No. 1, pp.83–95.

Yang, W., Wang, X., Tian, B., Xu, W. and Cheng, W. (2022) 'A multi-stage automatic evaluation system for sight-singing', *IEEE Transactions on Multimedia*, Vol. 25, pp.3881–3893.

Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y. and Feng, L. (2020) 'Deep attention based music genre classification', *Neurocomputing*, Vol. 372, pp.84–91.

Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., Xu, W., Lu, J., Hong, Z. and Wang, C. (2024) 'GTSinger: a global multi-technique singing corpus with realistic music scores for all singing tasks', *Advances in Neural Information Processing Systems*, Vol. 37, pp.1117–1140.

Zhou, F., Hang, R. and Liu, Q. (2020) 'Class-guided feature decoupling network for airborne image segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 3, pp.2245–2255.