ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Design and Implementation of an AI-Driven Virtual Choir Training Platform

Luanyu Zhao[1,*]

[1] Department of International Studies, Chugye University for Arts, Seodaemun-gu, Seoul, 03762, Korea

SUMMARY: *With rapid economic and technological development and the further enhancement of cultural soft power, the audience for choral art is gradually expanding. Ordinary citizens increasingly have opportunities to attend choral concerts, participate in choral competitions, and personally experience the charm of choral art. This study establishes a virtual choral training platform using virtual reality technology, creates a multimodal scene model, builds a comprehensive data transmission mechanism, and employs dot matrix technology to accurately simulate facial expressions. Voice analysis technology extracts vocal parameters of choral singing, including first formant, third formant, fundamental frequency, vocal range, and fundamental frequency perturbation. Subsequently, a Resnet-GRU-based choral singer recognition model is proposed, incorporating the embedded attention mechanism module SEnet to effectively enhance recognition performance. Results indicate that vocal patterns in singing exhibit greater discriminative power than those in natural speech. Compared to phoneme-based recognition, Resnet-GRU-based recognition achieves average accuracy improvements of 36.1%, 36.7%, and 29.8% for choral vocalization and retroflex consonants, respectively. Upon implementation, this method effectively enhances performers' emotional engagement while amplifying their expressive impact.*

KEYWORDS: *virtual reality; speech recognition; synthetic voice; ResNet-GRU; phonetic features*

1 Introduction

Choral singing is a collective art form, performed by groups ranging from as few as four to as many as hundreds of singers. Participants are divided into distinct vocal parts, working together to create artistic expressions [1]. In the digital age, people's demand for high-quality spiritual enrichment has grown increasingly strong, and enthusiasm for participating in cultural and artistic activities continues to rise. Against this backdrop, choral singing—as the most popular art form with the widest participation and relatively few constraints—is flourishing in a new era of diverse development [2, 3]. The virtual platform uses music training as its entry point, applying artificial intelligence and big data technologies to choral arts education and social music education. It aims to advance China's social music education by promoting the development of choral arts education [4]. The establishment of virtual platforms is based on multi-terminal interconnectivity, creating a visual training system and leveraging core technologies to provide music ensembles with specialized training models and comprehensive management systems [5]. New-type productive forces driven by artificial intelligence

*zhaoluanyu8@163.com
https://doi.org/10.65102/is2026644
Published: 30 April, 2026

accelerate innovation in music products, creation, presentation, and realization models, while simultaneously advancing music training and educational development [6, 7].

The application of artificial intelligence in the field of music encompasses two aspects. On one hand, it involves music composition, where AI's innovative potential has opened up entirely new perspectives for the diverse fusion of musical genres [8]. Chan et al. employed a machine learning-based automated music composition system to create music. This system generates musical sequences from musical fragments while satisfying various constraints, achieving an average satisfaction rating of 7.5 for the generated compositions [9]. Liu designed an AI-based automated music composition assistance system. In the generated musical works, adjacent notes exhibited an octave interval with an average proportion of 83.57%, demonstrating high scalability and service quality [10]. Chen proposed a deep music intelligent composition method, utilizing convolutional neural network theory for feature signal extraction from musical signals. The intelligent composition achieved an accuracy rate of 98%, with a feature signal frame loss rate below 5% [11]. Dai established an intelligent music teaching model through AI-powered technologies including smart perception, learning analysis, and affective computing. This approach enriched teaching methods while providing personalized practice services for students, enhancing music learning efficiency [12]. Doush et al. compared multiple intelligent algorithms during music creation and found the selected algorithms produced high-quality music. Among them, the neural network algorithm achieved an 83.3% accuracy rate in music generation, demonstrating the excellent performance of intelligent algorithms in music creation [13]. The melody generation model proposed by Zhao et al., based on recurrent neural networks and variational autoencoders, achieved favorable results in musical fluency, creativity, emotional expression, and harmony, offering novel approaches and perspectives for melody generation [14].

On the other hand, in AI-assisted music education, integrating artificial intelligence into the teaching system enables AI to assist instructors in explaining fundamental musical concepts and demonstrating techniques, leading to a significant overall improvement in training outcomes [15]. Zheng's research employed questionnaires and interviews to investigate the impact of AI technology on music education, revealing that AI helps students deepen their understanding of relevant theories, boosts learner confidence and interest, and enhances the quality of music instruction [16]. Zhang's research revealed that employing AI tools in vocal training influences singers' tonal consistency, vocal lightness, dynamic stability, melodic integrity, and linguistic imagery to varying degrees, ultimately achieving superior musical emotional expression [17]. Gai employed AI methods to study music education outcomes at the university level, finding that students achieved higher singing proficiency under AI intervention, with tonal beauty reaching greater heights as training duration increased [18]. Xu proposed a choral teaching assistance system utilizing music information retrieval and machine learning algorithms to analyze and compare recorded files, outputting visual charts matching user styles. Research revealed this system significantly boosted learners' willingness to engage in choral training [19]. Tianle built a music education cloud platform using technologies such as JOOQ, MongoDB, and MySQL. This platform enables mobile access, resource sharing, and granular management of music teaching materials, catering to students' personalized learning needs while enhancing teacher-student interaction [20]. Fan et al.’s study analyzed the impact of a virtual vocal education platform on students’ vocal skills. Empirical research demonstrated that training on this platform improved students' pitch and rhythm accuracy by an average of 15% and reduced task completion time by an average of 20% [21]. Qiusi's AI-based music education model achieved high satisfaction ratings among students, teachers, and parents. This model tangibly enhances students' musical proficiency and effectively expands the reach of

music education [22]. Liu integrated virtual reality and augmented reality technologies into music instruction. By simulating authentic performance scenarios and experiences, this approach stimulated students' learning interest, improved their musical perception and performance abilities, and elevated skill test scores to 83–99 points [23].

Artificial intelligence has established a cross-temporal and spatial platform for artistic skill exchange in music. However, research on choral music remains limited. Parametric visualization methods can deepen choir members' understanding of the conductor's intent. Simultaneously, constructing a tour rehearsal platform enhances artistic collaboration efficiency, providing technical support for building innovative pathways that integrate art and technology.

This paper constructs a highly realistic, multi-sensory virtual choir training system comprising facial capture, gesture capture, and speech recognition modules. Key data is preserved through multi-angle recording by virtual cameras. Within the speech recognition module, front-end processing for objective choir voice evaluation includes sampling, quantization, pre-emphasis, and windowed framing. Acoustic feature parameters are analyzed for meaning and extraction. The extracted choral speech signals are then fed into a ResNet network to extract spatial information, while a GRU network processes temporal information. An attention mechanism module (SEnet) is embedded during the convolutional processing, and a triadic loss function is employed during training to enable the network to recognize speech samples with higher similarity.

2 Design of an AI-Driven Virtual Choir Training Platform

2.1 Workflow Design

The development approach for virtual reality in vocal pedagogy should involve applied research across multiple domains, including 3D modeling, facial capture, gesture capture, speech recognition, and camera processing. Utilizing specific methodologies, one can visually analyze and evaluate changes in singers before and after virtual reality application, as illustrated in Figure 1.

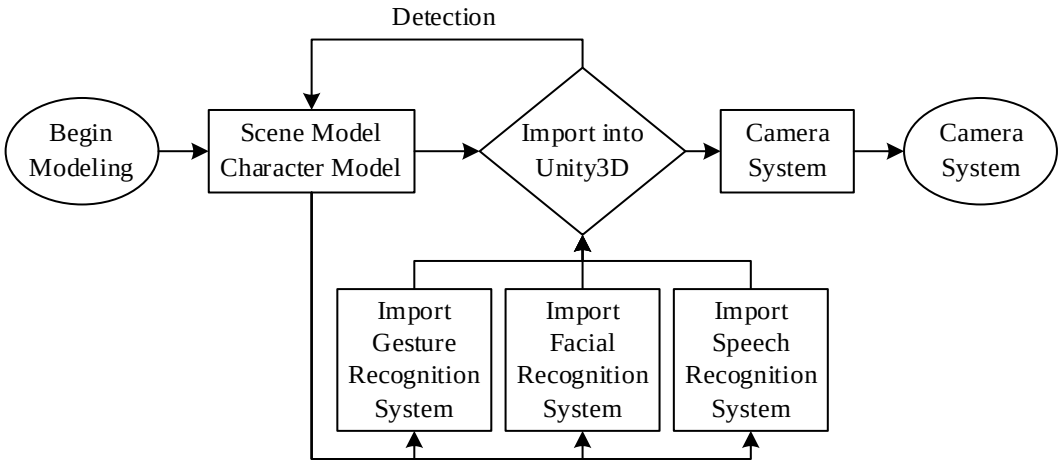


Figure 1: Workflow design

2.2 Virtual Model Creation

2.2.1 Scene Model

Taking a T-shaped stage scene as an example, its primary production method involves the following three steps:

(1) First, use spline editing in 3ds Max to create a T-shaped stage measuring 3000cm (length) × 2200cm (width) × 1000cm (height). Convert it to polygon editing and weld all vertices into a single entity to facilitate subsequent edge connection and wiring.

(2) Next, refine the model's details using mesh retopology. Employ tools like connect, extrude, bevel, and insert to refine specific structural elements. For the central stage floor, backdrop, and extended platform features, create these as separate models before bridging them to the main stage structure.

(3) Finally, create auxiliary models like the surrounding audience seating and overhead lighting rigs using geometric lofting combined with polygon editing. Primarily employ mirroring and duplicating modeling techniques during construction to gradually enrich the overall scene.

2.2.2 Role Model

Character models should be created under the box element. Utilize polygon editing to perform overall edge flow for the character's face and limbs, prioritizing quad-based edge flow. Focus edge flow on areas requiring virtual reality animation movement, such as facial features and limb joints, to refine structural relationships. For areas not involved in animation movement, effectively control the model's polygon count through techniques like collapsing edges and merging faces.

2.3 Key Module Design

2.3.1 Facial Capture Design

Due to the structured light principle employed by the system, light must be projected toward the face, and facial contours are determined by reading the illuminated surface data of the object. Therefore, when selecting facial capture devices, in addition to configuring distance sensors, microphones, and front-facing cameras, they must also feature sequentially arranged infrared lenses, flood illuminators, flood sensors, and dot matrix projectors.

Compared to methods that capture faces using two-dimensional images, The facial recognition accuracy of T_j at 0.1mm surpasses that of images (2), videos (1), and flat surfaces (0). Even under suboptimal lighting conditions R_i, the active facial information acquisition method—where the dot matrix projector emits light $-\sigma$ and receives light σ—does not compromise T_j's recognition efficiency. The optimization approach for its facial capture system can be modified as follows:

$$T_j(x,y)=\begin{cases}2, & if\ R_i(x,y)<-\sigma\\ 1, & if\ R_i(x,y)>\sigma\\ 0, & if\ -\sigma\le R_i(x,y)\le\sigma\end{cases} \tag{1}$$

2.3.2 Gesture Capture Design

Gesture capture is a technical challenge in virtual reality, requiring the singer to wear an Oculus

Quest 2 headset and connect hand-tracking devices to a computer. A depth-sensing camera is then mounted on the front of the headset, tilted downward at 13.4 degrees. This allows the singer to observe their hands in real-time during the VR experience and track their fingertip movements. The gestures, from left to right, are: backward, stop, and forward. If the fingertip position remains within the zero coordinate static zone (ZC), no movement is generated. When the fingertip extends forward beyond the static zone, the red progress bar indicating the subject's movement speed increases linearly with fingertip distance. When the finger joint moves in the opposite direction, palm-down toward the palm, and extends beyond the static zone, the red progress bar exhibits a subtle backward movement.

2.3.3 Design of the Speech Recognition Module

Speech recognition technology emerged in the 1950s and has evolved over nearly six decades. Numerous companies have since released voice application software. IBM's ViaVoice voice application software has matured significantly and now supports Chinese language recognition. IBM ViaVoice provides API functions for interfacing with other applications. This enables the development of voice recognition software with visual interfaces on PCs, utilizing IBM ViaVoice as the speech engine via these API functions. Through this voice recognition software, we can identify and extract useful voice information from speech, storing it in XML files. Figure 2 illustrates the voice information recognition process for this software.

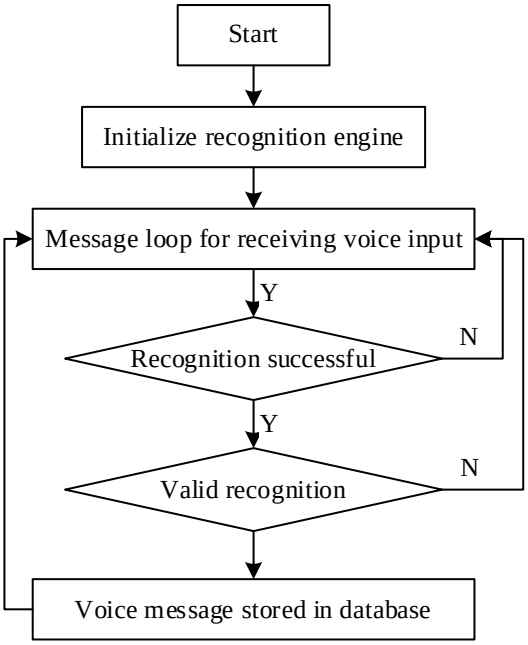


Figure 2: Voice information recognition process

2.3.4 Multi-View Camera Setup

To enhance the stability of virtual reality cameras and enable performers to review their overall performance—including movements and facial expressions—from multiple camera angles within the virtual space, necessary optimizations must be made to the underlying virtual reality software.

3 Voice Recognition-Based Choral Singer Identification and Training Model

3.1 Digitalization and Preprocessing of Choral Voices

3.1.1 Sampling and Quantization of Choral Voices

Choral voice audio samples are analog signals that must first be converted into digital signals for analysis. This involves sampling and quantization to obtain a signal with discrete time and amplitude values. Sampling is the discrete transformation of a continuous signal into a sequence of samples. According to the sampling theorem, when the sampling frequency is twice the signal bandwidth, no information is lost during sampling, and the original signal can be accurately reconstructed from the sampled signal.

Following sampling, the signal undergoes quantization—the discretization of the continuous waveform. Quantization involves dividing the entire amplitude range into discrete levels, ensuring samples within the same amplitude range share identical characteristics. The selection of quantization range and levels depends on the intended application of the digital representation.

3.1.2 Pre-Weighting of Choral Voices

Numerous factors contribute to signal attenuation, but the impact on high-frequency components within the oral cavity is significantly greater than that on low-frequency components. To facilitate analysis of the most pristine signal, pre-emphasis processing is applied to the vocal signal, enhancing its high-frequency components to approximate the original signal more closely. In experiments, pre-emphasis is typically achieved through filtering, commonly employing an FIR digital filter whose transfer function is expressed in Equation (2):

$$H(z) = 1 - \mu z^{-1} \tag{2}$$

Pre-emphasis serves three purposes: (1) It adds a zero to compensate for the reduced high-frequency components, bringing the resonant peak closer to its spectral position. This minimizes the residual effects of the vocal tract in artistic voices, ensuring the extracted features align more closely with the original vocal tract model. (2) The FIR digital filter acts as a high-pass filter, amplifying high-frequency components. (3) It also facilitates formant detection, reduces spectral instability, and enhances the stability of the quantized signal.

3.1.3 Window-Based Frame Segmentation for Choral Voices

Artistic voice lacks periodicity and is not a signal that changes according to regular patterns. Both its sampled values and characteristic parameters exhibit irregular variations over time. However, it is conventionally accepted that within an extremely short timeframe—typically 10 to 30 milliseconds—artistic voice can be regarded as a stable signal. Therefore, artistic voice can be characterized by two fundamental properties: time-varying nature and short-term stability. Based on these characteristics, artistic vocal signals are often segmented into distinct segments before processing, ensuring each segment maintains short-term stability. This operation is termed segmentation.

3.2 Vocal Parameter Extraction for Choral Voices

3.2.1 Resonance Peak Extraction

The upper resonant peaks determine the emotional coloration of the voice. All these characteristics collectively form the timbre of the voice. Many researchers worldwide focus on vocal quality or singer proficiency, often using the first and third formants as key evaluation metrics. Typically, the first formant largely determines the timbre of vowels in singing, while the third formant reveals the singer's individual characteristics and musical control.

Numerous methods exist for extracting formants, including bandpass filter combinations, spectral inversion, LPC detection, and AR modeling. After computational comparison, this experiment selected the AR method for extracting the first and third formants. This approach provides an excellent vocal tract model (provided the artistic voice samples contain minimal noise).

3.2.2 Fundamental Frequency Extraction

Fundamental frequency refers to the basic frequency produced when vocal cords vibrate, also known as the fundamental tone frequency. An excellent artistic voice should not only possess good timbre but also exhibit a certain degree of tension—that is, the ability to reach a certain pitch. Fundamental frequency is one of the key objective parameters for evaluating artistic vocal quality.

Numerous methods exist for extracting fundamental frequency, such as autocorrelation functions, average amplitude difference functions, autocorrelation, and wavelet transforms. After computational comparison, this experiment employs an improved algorithm to extract fundamental frequency periods, thereby calculating the fundamental frequency.

3.2.3 Pitch Range Extraction

Vocal range refers to the span of human voice production from low to high pitches. It can be categorized into physiological, conversational, and singing ranges, among others. The singing range is particularly common in artistic voice research, with key determinants being the characteristics of “multiple sound sources” and “multiple vibrations.” “Multiple sound sources” refers to a singer's ability to produce distinct vocal timbres, while “multiple vibrations” pertains to the vocal cords' capacity to modulate across various vibration frequencies. Singing range serves as a crucial indicator of artistic vocal quality, prompting extensive scholarly research in this area. A common method for estimating vocal range involves identifying the highest and lowest pitches within a song or its musical score. To mitigate data randomness and enhance accuracy, the mean and standard deviation are typically used as metrics.

3.2.4 Extraction of Fundamental Frequency Disturbances

Fundamental frequency perturbation is defined as the rate of change between the fundamental frequency of one acoustic wave cycle and that of the next cycle. Fundamental frequency perturbation is often used to measure the variation in acoustic wave values within the corresponding cycle, essentially reflecting the speed of vocal fold vibration between cycles.

The mathematical definition of fundamental frequency perturbation is given by Equation (3):

$$jitter = \frac{1}{N-1}\sum_{i=1}^{N}\left|1/F_{0i} - 1/F_{0(i-1)}\right| \tag{3}$$

In equation (3), $jitter$ is the mean value, N is the number of audio samples, and F_{0i} is the pitch of the i th note.

3.2.5 Extraction of Resonance Peak Perturbations

Resonance peak perturbation is defined as the rate of change between the resonance peak of one cycle and that of the next cycle. Resonance peak perturbation is often used to measure the variation in resonance peaks within corresponding cycles, essentially reflecting vocal quality or the technical proficiency of the singer.

Specifically, the first formant perturbation measures the rate of change of the first formant between adjacent cycles, while the third formant perturbation measures the rate of change of the third formant between adjacent cycles.

The mathematical definition of the first formant perturbation is given by Equation (4):

$$\frac{1}{N-1}\sum_{i=1}^{N}\left|1/F_{1i}-1/F_{1(i-1)}\right| \tag{4}$$

In equation (4), F_{1i} represents the first resonance peak of the i th cycle, and N denotes the number of audio samples.

The mathematical definition of the third resonance peak perturbation is given by equation (5):

$$\frac{1}{N-1}\sum_{i=1}^{N}\left|1/F_{3i}-1/F_{3(i-1)}\right| \tag{5}$$

In equation (5), F_{3i} represents the third resonance peak of the i th cycle, and N denotes the number of audio samples.

3.2.6 Average Energy Extraction

The definition of average energy represents the signal level under identical environmental conditions. Average energy is frequently used to measure the relative magnitude of vocal signals. The mathematical definition of average energy is given by Equation (6):

$$E_n=\sum_{k=-\infty}^{+\infty}x^2(k)w(n-k) \tag{6}$$

In equation (6), E_n represents the energy over a short time interval, $x(k)$ denotes the input signal, and $w(n-k)$ is the window function for the segment $n-k$.

3.3 ResNet-GRU-Based Choir Member Identification

3.3.1 Residual Network

The core concept of residual networks [24] is the identity map, which bridges convolutional layers. This cross-layer input allows for increased network depth without escalating model complexity. Moreover, due to this connection, the error gradient during backpropagation never falls below 1, enabling precise propagation of minor errors to preceding neural layers.

The forward propagation formula for residual networks is:

$$x_{k+1} = f\left(h(x_k) + F(x_k, W_k)\right) \tag{7}$$

Among these, x_k represents the input signal, x_{k+1} represents the output signal, and f represents the activation function. When it is an identity mapping, equation (8) becomes:

$$x_{k+1} = h(x_k) + F(x_k, W_k) \tag{8}$$

The forward propagation formula for the input signal from layer k to layer K is:

$$x_K = x_k + \sum_{i=k}^{K} F(x_i, W_i) \tag{9}$$

From equation (9), it can be seen that the input to layer K is the sum of the input to layer k and each residual block in between, constituting an additive computation. In contrast, traditional neural networks typically employ multiplicative operations. Compared to these, residual networks require significantly less computational effort.

The backpropagation formula for residual networks is:

$$\frac{\partial E}{\partial x_k} = \frac{\partial E}{\partial x_K}\left(1 + \frac{\partial}{\partial x_K}\sum_{k=1}^{K-k} F(x_i, W_i)\right) \tag{10}$$

As seen from Equation (10), during backpropagation of errors, applying the chain rule yields an additional identity term with a derivative of 1. This ensures that the gradient from layer K propagates stably to the preceding layer without encountering gradient vanishing issues.

3.3.2 Attention Mechanism Module

When convolving the feature parameters of speech signals, a large number of channels are generated, each containing information of varying importance. Therefore, this chapter introduces an attention mechanism module during the convolution process to assign greater weight to more significant information, thereby enhancing recognition performance. The attention mechanism module employed in the convolutional neural network in this chapter is SEnet [25]. This module enables the convolutional neural network to focus more on channels containing important information while reducing the number of network parameters.

The SE module primarily consists of two processes: Squeeze and Excitation. It takes feature maps as input, which are vectors of size $W \times H \times C$, where $W \times H$ represents the two-dimensional dimensions of the feature map, where C denotes the number of feature channels. The module compresses this into a $1 \times 1 \times C$ vector, then applies excitation processing. This process assigns varying weights to different feature channels based on their importance. The excited feature vector is then multiplied by the original feature map, yielding the final feature.

3.3.3 ResNet-GRU Model

Featuring parameters of the speech signal are input into the network for convolution operations. Each channel in the network represents a feature of the speech signal frame. Speech signals often contain noise and silent segments that cannot be completely removed even after preprocessing. Feature maps contain a significant amount of redundant information, and direct convolution can adversely affect recognition accuracy. Different feature channels contribute varying degrees to the final recognition outcome. To preserve valuable information while

minimizing redundant noise, the ResNet architecture incorporates the SEnet attention mechanism module. This module assigns distinct weights to feature channels, enabling the network to focus attention on meaningful speech segments while reducing the impact of noise and silence, thereby enhancing recognition performance. Simultaneously, GRU is employed to extract temporal features, thereby better leveraging contextual information within the speech signal.

3.3.4 Loss Function

When training networks using the traditional cross-entropy loss function, large datasets incur substantial computational overhead and exhibit slow convergence rates. Therefore, this section employs the triplet loss function [26] for network training. The triplet loss function demonstrates strong performance in recognizing similar samples, achieves faster convergence, and delivers superior recognition results in speaker identification tasks.

The triplet loss comprises three embedding vectors: an anchor vector, a positive vector, and a negative vector, collectively forming a triplet. The positive and negative vectors are referenced against the anchor vector. The positive vector originates from a different sample within the same category as the anchor vector, while the negative vector comes from a sample belonging to a different category than the anchor vector.

The mathematical expression is:

$$L = \max\left(\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha, 0 \right) \tag{11}$$

Here, α represents the boundary value. The primary purpose of introducing the boundary value definition is to prevent the model from taking shortcuts—specifically, training the distance between positive and negative vectors to match that of the anchor vector. This would cause the discriminative power between positive and negative vectors to disappear. If α is too large, it causes excessive model loss, making network convergence difficult or even impossible. If α is too small, distinguishing positive and negative vectors becomes challenging.

The original ternary loss function measures similarity using Euclidean distance. This section adopts cosine similarity as the metric, making the normalized similarity in the actual validation phase more convenient and intuitive, and yielding better results in speech signal processing.

Similar to the fundamental principle of the ternary loss function based on Euclidean distance, the cosine similarity-based ternary loss function also aims to bring positive vectors closer to the anchor vector while pushing negative vectors farther away. Thus, Equation (11) can be rewritten as:

$$L = \frac{1}{N} \sum_{i=1}^{N} \max(s_i^{a,p} - s_i^{a,n} + \alpha, 0) \tag{12}$$

Here, $s_i^{a,p}$ denotes the cosine similarity between the anchor vector and the positive vector, $s_i^{a,n}$ denotes the cosine similarity between the anchor vector and the negative vector, and α is the threshold value.

The operation principle of the cosine similarity-based ternary loss function adopted in this section is as follows: First, a speech signal is randomly selected as the anchor vector. Then, a positive vector is randomly chosen from the same speaker's speech, and a negative vector is randomly selected from other speakers' speech.

4 Virtual Choir Training System Testing

4.1 Distinctiveness of Acoustic Features in Choral Acoustics

4.1.1 Distinctiveness of Pitch

First, a subset of speech samples was randomly extracted from the RSS corpus based on the following principles: two speakers of the same gender; speech samples representing both normal speech and singing modes for the target speaker; selected speech samples originating from the same text source but differing in vocal mode; each sample lasting approximately 3 seconds. Pitch information within each recording was represented using NCCF coefficients, as illustrated in Figures 3 and 4.

In Figure 3, “spk1 reading” and “spk1 singing” represent the results for the first speaker's normal speech and singing modes, respectively. Similarly, “spk2 reading” and “spk2 singing” in Figure 4 denote the corresponding results for the second speaker. Figure 3 displays the NCCF coefficient feature statistics for speech signals corresponding to normal speaking and singing modes by the same female speaker using identical text. Similarly, Figure 4 compares feature distributions between different speakers, showing NCCF coefficient statistics for normal speaking and singing modes between two distinct female speakers.

In Figure 3, it is evident that under identical speaker and text conditions, the NCCF trajectory of normal speech deviates from that of singing. Comparing the NCCF trajectories in Figure 4 reveals that the difference between singing and singing vocalization across two distinct speakers is greater than the difference in reading vocalization, even under text-related tasks. Combining Figures 3 and 4, preliminary conclusions drawn from the NCCF feature space indicate: - The NCCF feature space differs between singing and normal speech modes for the same speaker; - Differences in singing speech are greater between different speakers than differences in normal speech.

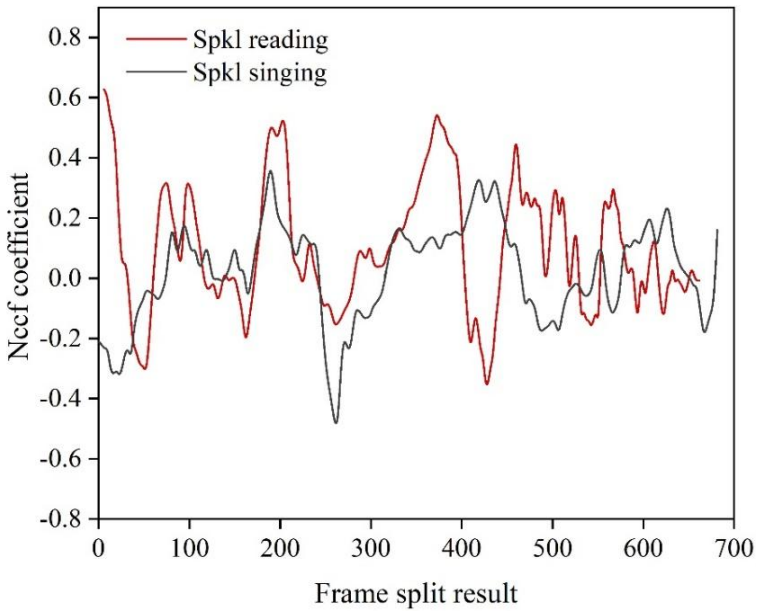
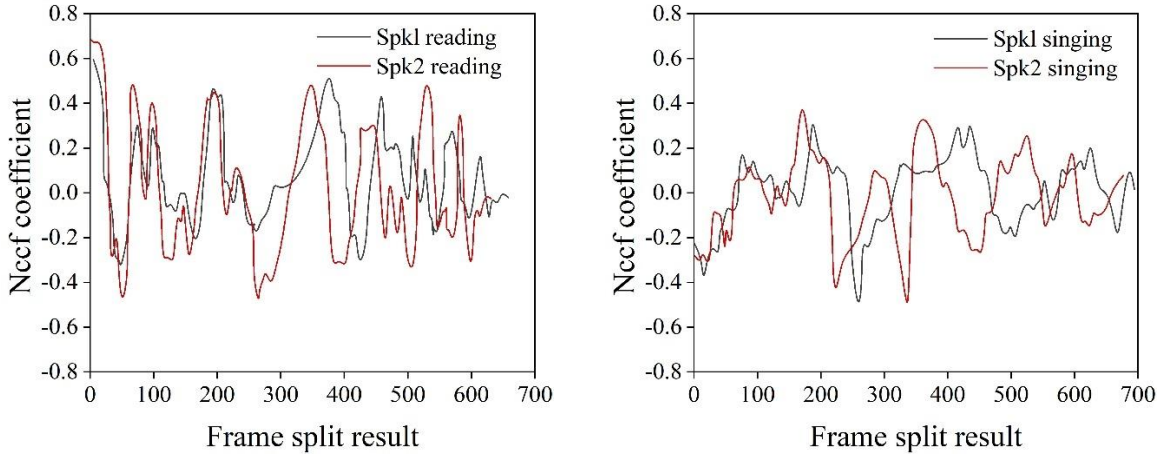


Figure 3: The NCFF coefficients of singing and speech used by the same speaker



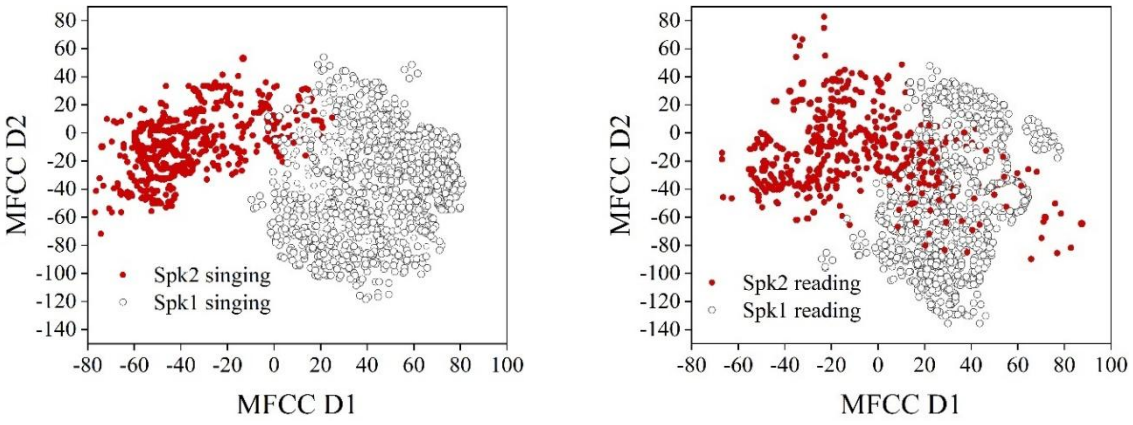
(a) Speech (b) Sing

Figure 4: The NCCF coefficients of speech and singing are better for different people

4.1.2 Discriminative Power of MFCCs

The speech samples extracted in this section share identical conditions with those used in the previous section. MFCC has been widely adopted for voiceprint recognition. First, 20-dimensional (c0-c19) MFCC features were extracted. Subsequently, to visualize and analyze MFCC discrimination between different speakers more clearly, the Principal Component Analysis (PCA) algorithm was applied to reduce the 20-dimensional MFCC to a two-dimensional feature space. Finally, feature analysis was conducted within this low-dimensional space. “spk1 singing,” “spk2 singing,” “spk1 reading,” and “spk2 reading” correspond exactly to the descriptions in the previous section. The horizontal axis of both subfigures represents the magnitude of the first dimension coefficient in the two-dimensional space, while the vertical axis represents the magnitude of the second dimension coefficient. Each point in the distribution represents one frame of speech data.

Figure 5 illustrates the two-dimensional MFCC distributions of normal speech and singing for two distinct female speakers using identical text. The overlap between red dots and black circles in the left subplot is observed to be smaller than in the right subplot. This indicates that singing speech exhibits greater discriminative power in the MFCC feature space compared to normal speech. By comparing the same-colored regions in the left and right subplots, significant differences exist between singing and normal speech in the MFCC feature space, even for the same speaker with identical text. Therefore, it is speculated that using singing voice data rather than normal speech data to characterize speaker identity may yield better results. In practice, even when discriminative information is observed in the feature space, it does not guarantee that the speaker model will effectively utilize this discriminative power.



(a) Speech (b) Sing

Figure 5: Distribution of PCA features and speech comparison

4.2 Experimental Design

4.2.1 Recording of Experimental Content

Recording sessions may take place in a studio or concert hall, utilizing software audio workstations such as Pro Tools or Cubase as the multitrack recording platform. Two identical large-diaphragm condenser microphones (e.g., Neumann U87) are employed, arranged in an XY stereo configuration to capture greater dynamic range and a wider soundstage. Each vocal part may be recorded by 4-6 choir members. Multiple takes are layered and doubled to create a choral ensemble effect. An equalizer (EQ) is used to balance the choral timbre, a compressor to control overall dynamic range, and a reverb unit to simulate spatial depth. After recording, individual vocal parts are exported as WAV audio files for use in experimental project development.

4.2.2 Experimental Platform

The processor model for the hardware platform configuration in this document is the Intel® Core™ i7-7700K CPU @ 4.20GHz. The GPU is an NVIDIA GeForce GTX 1080. The development environment PyCharm is installed on the Windows 10 system. The program code runs in environments created by Anaconda based on Python 2.7 and Python 3.7. The network model is implemented using TensorFlow 1.5.0 and Keras 2.1.4. TensorFlow is a comprehensive software library open-sourced in 2015, providing diverse algorithmic models for machine learning. In this study, audio feature extraction utilizes the Madmom computational toolkit to extract Log-mel acoustic features.

For experiments, the training and testing datasets comprise 90% and 10% of the corpus, respectively. Cross-validation is employed during model training, where one portion of the data serves as the training set while another forms the validation set. To mitigate detection randomness, each model undergoes five independent training iterations, with the average recognition rate serving as the performance metric.

4.3 Analysis of Experimental Results

4.3.1 Comparison of Phoneme-Based Recognition and Articulation Feature-Based Recognition

After feature extraction and normalization, this paper will train the ResNet-GRU model using

the obtained feature dataset. Precision, recall, and F-score are selected as evaluation metrics. To obtain an optimized choral vocal recognition model, comparative experiments are conducted on the following recognition methods and model parameters: class weights, window size, number of convolutional kernels, random dropout rate, and phoneme recognition. The experimental results are shown in Table 1.

As shown in the table, the recognition accuracies based on phonemes are: 51.2%, 56.3%, 53.4%, 42.4%, 20.5%, 28.1%, 48.7%, 43.4%, 64.2%, and 77.6%. For certain phonemes with smaller sample sizes, recognition rates were low—e.g., phoneme q achieved only 20.5%. The average recognition accuracy for retroflex phonemes based on phoneme-based models was 53.5%, 25.8%, and 53.4%. In contrast, the ResNet-GRU-based model achieved average recognition rates of 89.6%, 62.5%, and 83.2% for choral vocal phonation, demonstrating significant improvements across all metrics. The underlying reason lies in the improved sample balance achieved by the Resnet-GRU-based recognition model. Experimental results indicate that choral vocalization and retroflex sounds exhibit significant acoustic feature differences, enabling effective recognition through classification of these phonetic characteristics.

Table 1: The evaluation of experimental results of different model structures and parameters

Model structure and parameters	Dental B1			palatal sound B2			Retroflexion B3			Other B4	Mean A
	F1-score (P,R)(%)			F1-score (P,R)(%)			F1-score (P,R)(%)			F1-score (P,R)(%)	F1-score (P,R)(%)
The weight of the irrelevant											
Random discard rate:0.34	36.5			33.2			48.3			70.8	36.5
window size:15×80bins	(25.6,60.1)			(20.8,72.7)			(61.4,41.5)			(93.5,59.3)	(25.6,60.1)
Number of convolution kernels:450											
The weight of the irrelevant											
Random discard rate:0.6	40.2			33.5			46.2			86.2	40.2
window size:15×80bins	(34.2,46.8)			(21.6,88.5)			(34.2,47.4)			(93.6,78.2)	(34.2,46.8)
Number of convolution kernels:450											
The weight of the irrelevant											
Random discard rate:0.6	35.6			31.5			49.7			73.2	36.7
window size:15×80bins	(26.3,53.7)			(21.4,88.4)			(34.2,46.7)			(90.2,60.5)	(26.4,54.6)
Number of convolution kernels:171											
The weight of the irrelevant											
Random discard rate:0.34	89.6			62.5			83.2			95.7	90.1
window size:15×80bins	(96.4,83.5)			(55.4,73.3)			(83.6,81.5)			(93.7,95.3)	(89.2,89.6)
Number of convolution kernels:450											
The weight of the irrelevant											
Random discard rate:0.32	75.2			52.6			72.8			88.9	76.4
window size:15×80bins	(73,4,75.8)			(54.5,89.6)			(75.7,70.1)			(94.5,88.7)	(72.4,80.7)
Number of convolution kernels:448											
Choir pronunciation	z	c	s	j	q	x	zh	ch	sh		
Voiceprint recognition	51.2	56.3	53.4	42.4	20.5	28.1	48.7	43.4	64.2	77.6	62.7
	53.5			25.8			53.4			77.6	

4.3.2 Confusion Matrix of Detection Models

To evaluate the model's recognition performance, we fed the test set from the corpus into the model and calculated the confusion matrix comparing predicted results with actual results, as shown in Figure 6. As shown, the numbers in the confusion matrix represent the quantity of different phonetic features, with darker colors indicating a higher number of features. The numbers along the diagonal represent the quantity of phonetic features predicted correctly by the model, while the numbers in other cells represent the quantity predicted incorrectly. The number in the upper-right corner is 1309, appearing darkest because it represents the largest proportion of other consonants. Overall, the diagonal elements of the confusion matrix are relatively high, indicating a high prediction accuracy rate for the model.

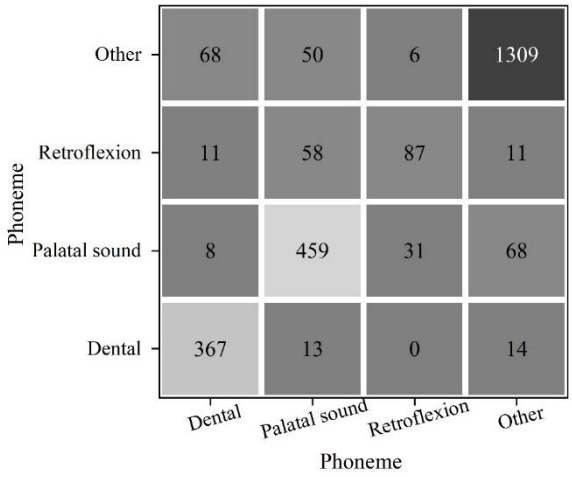
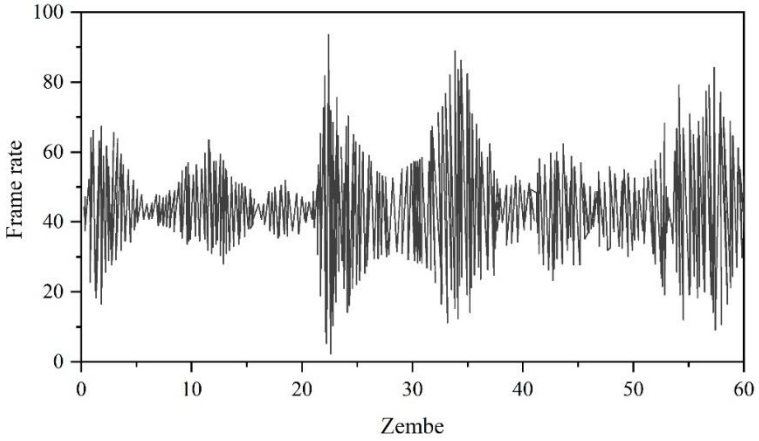


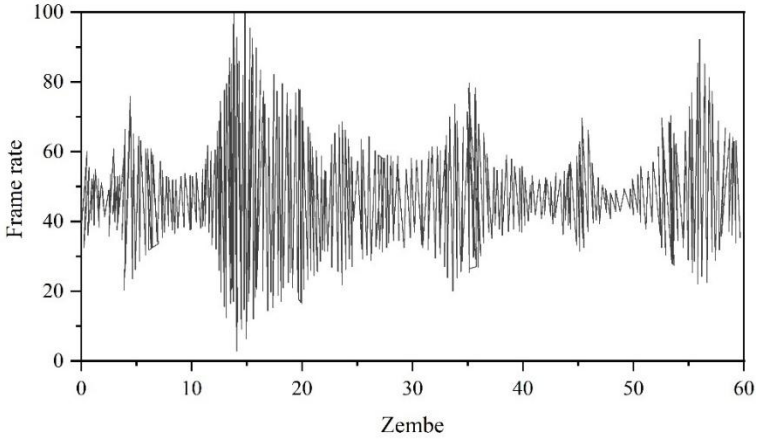
Figure 6: Confusion matrix thickness of choral voice detection

4.3.3 Instance Recognition Analysis

Based on the trained model, this paper processes the choir audio by extracting features and feeding it into the aforementioned recognition model to perform frame-level choir vocal recognition. The choir audio features the song “Great Love Without Borders.” Figure 7 displays partial recognition results for sung syllables. The figure indicates that the model achieved overall successful recognition, with only a small number of frames containing errors. However, the model demonstrates strong overall recognition performance, providing effective feedback on the phonetic classification of choral vocal performances.



(a) Previous performance



(b) The latter part of the performance

Figure 7: Chorus vocal pronunciation characteristics and model recognition results

4.4 Analysis of System Application Effectiveness

The recognition system was subsequently applied to a choral training platform. Sixteen music majors were selected to perform a song in two groups: Group A employed traditional methods, while Group B utilized the method described herein. Using a peer evaluation scale, five experts scored participants' a cappella performance recordings on a 10-point scale.

The scoring rubric comprised seven dimensions: Song Singing Coherence (S1), Breath Control (S2), Resonance Utilization (S3), Pitch and Rhythm Accuracy (S4), Articulation (S5), Musicality (S6), and Emotional Expression (S7). For each dimension, the average of the five scores per participant across different emotional states was calculated. Statistical analysis was then performed on the average scores of all 16 participants across both experimental groups, with results shown in Table 2. Significant differences ($p<0.01$) were observed between the two groups in five dimensions (excluding pitch and rhythm, and enunciation) across all three emotional states, with Group B consistently outperforming Group A. The greater coherence in singing by Group B may be attributed to the sequence of trials, where Group B performed after Group A. Group B demonstrated significant improvements in breath control, resonance, and musicality, alongside richer emotional expression. These findings indicate that the ResNet-GRU-based virtual choir training platform positively enhances emotional engagement in vocal performance among participants.

Table 2: Vocal other assessment scale and significance level

Dimension	Negative direction			Neutral			Forward direction		
	A group	B group	p	A group	B group	p	A group	B group	p
S1	6.713	6.832	<0.01	6.729	7.041	<0.01	6.843	7.109	<0.01
S2	6.402	6.995	<0.01	6.673	7.129	<0.01	6.605	7.084	<0.01
S3	6.351	6.962	<0.01	6.573	7.415	<0.01	6.578	7.425	<0.01
S4	6.188	6.285	0.22	6.304	6.391	0.26	6.381	6.473	0.24
S5	6.713	6.598	0.09	6.643	6.620	0.47	6.694	6.713	0.44
S6	6.594	7.359	<0.01	6.897	7.396	<0.01	6.711	7.627	<0.01
S7	6.629	7.884	<0.01	6.982	7.734	<0.01	6.729	7.794	<0.01

5 Conclusion

This paper establishes a choir training platform using virtual reality technology, implementing sampling, quantification, pre-emphasis, and windowed frame segmentation for the objective evaluation of choral voices. It analyzes the meaning and extraction of acoustic feature parameters. Subsequently, a choir singer recognition model based on ResNet and GRU was developed and validated through experimental design. The following conclusions were drawn:

(1) Comparing the distribution distinctiveness of singing and normal speech across different acoustic feature spaces—specifically pitch and MFCC parameters—revealed that singing exhibits superior discriminative capability.

(2) Based on the established choir vocal corpus, analysis revealed that the average recognition accuracy for phoneme-based choir vocal production and retroflex consonants was 53.5%, 25.8%, and 53.4%, respectively. In contrast, the average recognition rates for choral vocalization based on articulatory features reached 89.6%, 62.5%, and 83.2%, representing improvements of 36.1%, 36.7%, and 29.8%, respectively.

(3) Expert evaluations using a vocal assessment rubric revealed statistically significant differences in emotional expression scores between participants using the ResNet-GRU-based virtual choir training platform and those using traditional methods. Performers utilizing the ResNet-GRU platform also demonstrated notable improvements in breath control, musicality, and resonance.

References

[1] Rolsten, K. (2016). The production of quality choral performance: A review of literature. Update: Applications of Research in Music Education, 35(1), 66-73.

[2] Redden-Liotta, C. J. (2012). WWWhere did you find that piece? Finding repertoire in the digital age: A guide to online resources for choral directors. Choral Journal, 52(7), 47-51.

[3] Campbell, P. A. (2022). Chorus and Crisis in the Contemporary United States. In Staging 21st Century Tragedies (pp. 71-77). Routledge.

[4] Leonard, J., Cadoz, C., Castagné, N., Florens, J. L., & Luciani, A. (2013, October). A virtual reality platform for musical creation: GENESIS-RT. In International Symposium on Computer Music Multidisciplinary Research (pp. 346-371). Cham: Springer International Publishing.

[5] Tang, Y., & Zeng, X. (2024). Research on vocal music online educational platform based on internet platform. International Journal of Web Engineering and Technology, 19(4), 360-378.

[6] Mycka, J., & Mańdziuk, J. (2025). Artificial intelligence in music: recent trends and challenges. Neural Computing and Applications, 37(2), 801-839.

[7] Holland, S. (2013). Artificial intelligence in music education: A critical review. Readings in music and artificial intelligence, 239-274.

[8] Hernandez-Olivan, C., & Beltran, J. R. (2022). Music composition with deep learning: A review. Advances in speech and music technology: computational aspects and

applications, 25-50.

[9] Chan, M., Potter, J., & Schubert, E. (2006, August). Improving algorithmic music composition with machine learning. In Proceedings of the 9th International Conference on Music Perception and Cognition, ICMPC.

[10] Liu, Anna. "Multi-genre Digital Music Based on Artificial Intelligence Automation Assisted Composition System." Informatica 48.5 (2024).

[11] Chen, J. (2022). Construction of Music Intelligent Creation Model Based on Convolutional Neural Network. Computational Intelligence and Neuroscience, 2022(1), 2854066.

[12] Dai, D. D. (2021). Artificial intelligence technology assisted music teaching design. Scientific programming, 2021(1), 9141339.

[13] Doush, I. A., & Sawalha, A. (2020). Automatic music composition using genetic algorithm and artificial neural networks. Malaysian Journal of Computer Science, 33(1), 35-51.

[14] Zhao, H., Min, S., Fang, J., & Bian, S. (2025). AI-driven music composition: Melody generation using Recurrent Neural Networks and Variational Autoencoders. Alexandria Engineering Journal, 120, 258-270.

[15] Merchán Sánchez-Jara, J. F., González Gutiérrez, S., Cruz Rodríguez, J., & Syroyid Syroyid, B. (2024). Artificial intelligence-assisted music education: A critical synthesis of challenges and opportunities. Education Sciences, 14(11), 1171.

[16] Zheng, H., & Dai, D. (2022). Construction and Optimization of Artificial Intelligence-Assisted Interactive College Music Performance Teaching System. Scientific programming, 2022(1), 3199860.

[17] Zhang, Y. (2025). Increasing Emotional Perception in Academic Singing During Vocal Performance: The Use of AI Solutions. International Journal of Human–Computer Interaction, 1-9.

[18] Gai, M. (2025). Implementation of an Innovative Approach to Vocal Training in College: The Case of Artificial Intelligence Technologies NSynth, Sing Like Me, and Flow Machines. The Asia-Pacific Education Researcher, 1-11.

[19] Xu, Y. (2024, December). Applying Artificial Intelligence Techniques to Build a Singing Evaluation System in Choral Singing. In Proceedings of the 2024 2nd International Conference on Information Education and Artificial Intelligence (pp. 469-475).

[20] Tianle, Z. (2020, February). Research on the Construction of University Music Teaching Cloud Platform for Mobile Terminals. In 2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) (pp. 109-113). IEEE.

[21] Fan, X., Chen, R., & Zheng, J. (2025). Research on the Teaching Value and Promotion of Virtual Simulation Platform for Digital Human Vocal Music Education. Frontiers in Interdisciplinary Educational Methodology, 2(2), 108-123.

[22] Qiusi, M. (2022). Research on the improvement method of music education level under the background of AI technology. Mobile information systems, 2022(1), 7616619.

[23] Liu, J. (2025, February). Application of Artificial Intelligence Technology to Assist Music Teaching in University Classrooms. In Proceedings of the 2025 International Conference on Big Data and Informatization Education (pp. 149-154).

[24] Xiaoshuang Lv,Xin Ma,Wei Peng,Ke Li & Chengdong Li. (2025). A novel mechanism-guided residual network for accurate modelling of scroll expander under noisy and sparse data conditions. Complex & Intelligent Systems,11(10),418-418.

[25] Jia Zhen,Wang Kai,Li Yang,Liu Zhenbao,Qin Jian & Yang Qiqi. (2022). High Precision Feature Fast Extraction Strategy for Aircraft Attitude Sensor Fault Based on RepVGG and SENet Attention Mechanism. Sensors,22(24),9662-9662.

[26] Wu Hanxiao,Shen Fei,Zhu Jianqing,Zeng Huanqiang,Zhu Xiaobin & Lei Zhen. (2022). A sample-proxy dual triplet loss function for object re-identification. IET Image Processing,16(14),3781-3789.