ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

A strategy for extracting and evaluating timbre performance features of singing skills in music education based on music waveform spectral modeling

Wenjing Wu[1,*] *and Jacquline Tham*[2]

[1] Kaifeng University, Dongjing Avenue, 475000, Kaifeng City, Henan Province, China.
[2] Postgraduate Centre, Management and Science University, 40100, Shah Alam, Selangor Darul Ehsan, Malaysia

SUMMARY: *In the process of music education, extracting timbre characteristics from a singer's voice is an indispensable step for grasping musical style and enhancing vocal performance skills. This paper first proposes a method for extracting timbre characteristics in vocal techniques for music education. It employs autocorrelation and MFCC methods to extract fundamental frequency for modeling musical waveform spectra. Subsequently, the proposed timbre extraction method is validated through experiments analyzing vocal timbre characteristics and error analysis. Furthermore, a scoring method for evaluating singers' performance levels is presented and experimentally tested. Results demonstrate that the proposed evaluation criteria exhibit the highest consistency with expert assessments, achieving correlation coefficients of 0.7826 and 0.7687. Both values surpass existing pitch/rhythm evaluation metrics, thereby validating the effectiveness of this methodology.*

KEYWORDS: *Autocorrelation Method; MFCC; Feature Extraction; Scoring Method*

1 Introduction

The spectral shape of a music waveform refers to the distribution characteristics of its short-time Fourier transform spectrum [1]. During audio analysis, many features are not readily apparent in the time domain and require transformation into the frequency domain for analysis [2]. The Fourier transform serves as a crucial tool for spectral analysis, linking the time domain and frequency domain of a signal to enable effective extraction of information across both domains [3]. In practice, signals processed by computers are discrete. To accommodate the short-term stationarity of audio signals, the short-time Fourier transform (STFT) was developed from the Fourier transform specifically for audio analysis [4]. Timbre is typically defined as a measure listeners use to distinguish different sounds and instruments at the same loudness level [5]. Timbre is primarily determined by frequency-domain statistical features, and timbre feature extraction essentially involves capturing these frequency-domain characteristics [6]. Typically, dozens or hundreds of features are used to describe an entire song, significantly reducing the total data volume to be processed. This approach eliminates redundant information irrelevant to music analysis tasks while transforming raw data into more suitable representations [7].

Extraction of timbre-related features is crucial for instrument recognition, prompting numerous studies focused on timbre extraction for this purpose [8]. Reference [9] indicates that

*kfviolin2023@163.com
https://doi.org/10.65102/is20261077 Published: 10 June, 2026

the Fast Fourier Transform (FFT) can capture the musical spectra of different instruments. By combining specific timbre features derived from single-instrument recordings with multidimensional unsupervised machine learning techniques for timbre feature recognition, instruments can be accurately classified and musical timbre quality evaluated. Reference [10] employed a method where listeners were exposed to timbres varying with pitch to cultivate their ability to distinguish instrument sources from sounds of different pitches. Results showed listeners could identify instruments based on the most relevant timbral constancy and associated recognition experience. Furthermore, integrating musical spectrograms enhanced listeners' recognition rates. Reference [11] applied harmonic time clustering to segment instrumental audio signals into acoustic events for timbre and quality recognition tasks. Feature reduction and support vector machines were then employed to accurately identify instruments such as brass and strings. Reference [12] employed learning vector quantization neural network learning and short-time Fourier transform techniques to extract acoustic components like timbre from music sources. By reducing the dimensionality of classifier feature vectors, model training and recognition efficiency were improved. At a feature dimension of 24, the model achieved an 81.2% weighted recognition accuracy for musical instruments. Reference [13] reviewed the current state of research on instrument recognition and designed a neural network-based instrument recognition model. The model demonstrated outstanding performance on both training and validation sets, achieving recognition accuracies ranging from 0.86 to 0.99 for instruments such as the guitar. Reference [14] distinguishes audio types of different timbres and instruments by analyzing timbral similarity, with geometric distance calculations playing a crucial role. They constructed a 7-dimensional timbre space composed of multiple databases and employed machine learning algorithms to compute Euclidean distances within this space for studying timbral similarity.

To enhance the effectiveness of intelligent analysis of vocal techniques in music education, Reference [15] employed music waveform feature extraction technology to conduct intelligent analysis of vocal techniques, achieving favorable results. Subsequently, an optimal waveform selection algorithm was used to analyze the model's tracking performance of vocal techniques and variations in transmission parameters. A multi-method integration strategy improved both vocal performance skills and music teaching outcomes. Reference [16] designed an audio-extraction-based vocal teaching method. This method achieves an average recognition time of 0.006 seconds for pharyngeal training, effectively extracting pharyngeal training features during singing to enable pharyngeal protection training in vocal learning. Meanwhile, the piano audio synthesis model proposed in [17] enables timbre modification and generates piano-specific tonal characteristics. The piano timbre library generation system designed based on this model flexibly produces piano timbre libraries of varying quality, achieving superior teaching outcomes in piano instruction with its assistance.

Timbre perception is a crucial element for rendering emotion and shaping expressiveness in musical works. Timbre evaluation can further clarify the learning outcomes of vocal techniques in music education [18]. Reference [19] details specific experimental procedures and analysis processes to obtain timbre perception feature values for various timbre materials. The timbre perception model established using support vector regression effectively predicts different timbre perception features. Reference [20] constructs a machine learning-based automatic evaluation model to assess musicians' timbre capabilities in trumpet performance. Using human evaluation results as a benchmark, it identifies spectral peak stability as the key timbre factor influencing trumpet sound quality. Reference [21] performed preprocessing on timbre signals in an online vocal teaching system—including removing silent segments, pre-emphasis, and windowing—to extract timbre signal features. A timbre signal feature comparison model was constructed to reduce the dimensionality of the feature vector. Based on

the SAGA algorithm, a timbre evaluation model for online vocal teaching was developed, and experiments demonstrated the feasibility of the proposed method.

Individual differences in timbre perception hold promise as a future research direction. This not only facilitates a more comprehensive exploration of cognition and possibilities in music and interdisciplinary fields but also provides stronger theoretical support for practical applications such as music composition and music education [22, 23].

The paper first introduces methods for audio signal preprocessing, focusing on windowing and frame segmentation as well as endpoint detection. It then presents specific theoretical steps for extracting vocal feature parameters and commonly used techniques in practical extraction. The fundamental frequency is extracted using the autocorrelation fundamental detection algorithm (ACF). Mel-frequency cepstral coefficients are obtained through Mel-frequency cepstral coefficients and Fourier transform. Experiments analyze the spectra of pitched notes in male and female voices, observing acoustic features in spectral plots and comparing differences between notes of varying pitches. Further analysis examines patterns in consonant duration, vowel pitch variation, and acoustic statistics of vibrato. Error analysis is conducted on the note onset detection module. Finally, the paper proposes an evaluation method for assessing singers' performance levels and conducts subjective evaluations of vocal quality.

2 Music Signal Preprocessing

2.1 Window-Based Frame Segmentation

The frame length and frame shift in windowed framing are illustrated in Figure 1. This design ensures smooth transitions between frames, maintaining continuity. The overlapping portion between the preceding and succeeding frames is termed the frame shift, with the ratio of frame shift to frame length typically ranging from 0 to 0.5. Frame segmentation is achieved by applying a movable, finite-length window for weighting, thereby generating a windowed music signal.

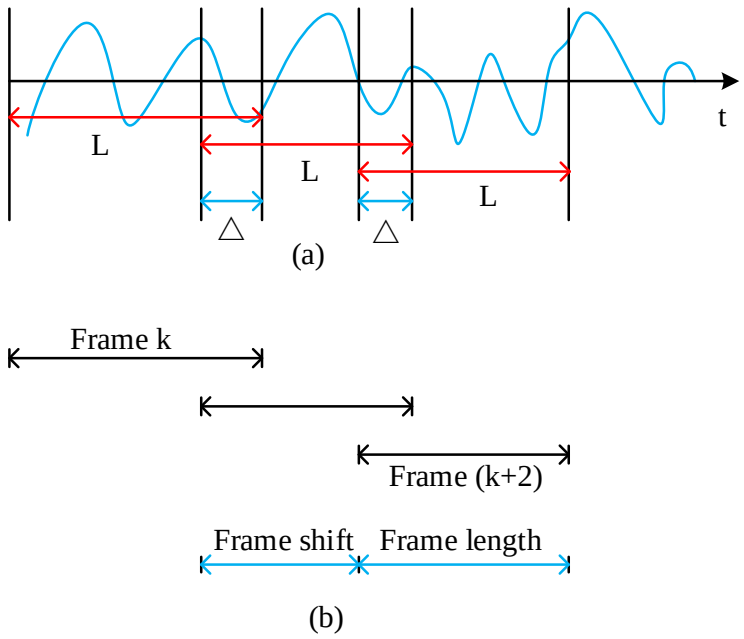


Figure 1: Frame length and frame shift in windowed framing

Common window functions used in digital signal processing include the rectangular window and the Hamming window, whose expressions are as follows (where N is the frame length):

Rectangular window:

$$w(n) = \begin{pmatrix} 1, & 0 \le n \le (N-1) \\ 0, & n = \text{Other values} \end{pmatrix} \tag{1}$$

Han Ming Window:

$$w(n) = \begin{pmatrix} 0.54 - 0.46\cos[2\pi n/(N-1)] & 0 \le n \le N-1 \\ 0, & n = \text{Other values} \end{pmatrix} \tag{2}$$

2.2 Endpoint Detection

Endpoint detection technology plays a crucial role in speech signal processing, enabling the precise identification of the onset and offset points within a speech signal to distinguish between speech and non-speech signals. When analyzing musical signals, such as plucking a single note, determining its start and end points is essential to observe the trajectory of its sustained tonal variation. Endpoint detection constitutes a vast research domain, and this paper provides a brief overview of its commonly used methods.

(1) Short-Term Energy Analysis

Speech is typically divided into silent segments, voiceless segments, and voiced segments. Based on the inherent characteristics of speech signals, within a short timeframe of 10–30 ms, it can be regarded as a quasi-steady-state process exhibiting short-term properties. Let the short-term energy spectrum E_n of the speech signal $x_n(m)$ in the nth frame be denoted as, then its calculation formula is as follows:

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \tag{3}$$

In the formula, N represents the frame length.

(2) Short-Term Zero-Crossing Rate Analysis

The short-term zero-crossing rate indicates the number of times the speech signal waveform crosses the horizontal axis within a frame. It can be used to distinguish between voiceless and voiced sounds, as the high-frequency range of speech signals exhibits a high zero-crossing rate, while the low-frequency range has a lower rate. For continuous signals, a zero crossing occurs when the waveform crosses the time axis. For discrete signals, a zero crossing occurs when the sign of the value at an adjacent sample point changes. The zero crossing rate is the number of times the sign of a sample point changes. The short-term zero crossing rate Z_n of speech signal $x_n(m)$ is defined as:

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1} |\operatorname{sgn}[x_n(m)] - \operatorname{sgn}[x_n(m-1)]| \tag{4}$$

In the formula, sgn[] is the sign function, that is:

$$\mathrm{sgn}[x] = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \tag{5}$$

By combining these two parameters, the effective speech components in the original signal can be identified.

3 Method for Extracting Timbre Expression Characteristics in Vocal Techniques for Music Education

3.1 Extracting Fundamental Frequency Using the Autocorrelation Method

The Auto-Correlation Fundamental Detection Algorithm (ACF) is a fundamental detection method based on the theory of speech signal time-domain analysis. The fundamental principle of the auto-correlation function method involves multiplying the speech signal by an auto-correlation function derived from the speech signal itself to generate a new speech signal function. This function produces a maximum value at a time delay corresponding to the fundamental period. Leveraging this property, we estimate the fundamental frequency of a speech signal by locating the maximum point of its autocorrelation function.

Generally, speech signals are categorized into two types: discrete digital speech signals and random or periodic signal sequences.

The autocorrelation function for a discrete digital speech signal sequence is defined as follows:

$$R(k) = \sum_{-\infty}^{\infty} x(m) * x(m+k) \tag{6}$$

Here, $x(m)$ denotes the digital speech signal function, and k represents the signal delay point number.

We define the autocorrelation function of a random signal sequence or periodic signal sequence as:

$$R(k) = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{-\infty}^{\infty} x(m) * x(m+k) \tag{7}$$

The autocorrelation function exhibits the following properties: If the digital speech signal function $x(m)$ is a periodic function with period T, then its autocorrelation function is also a periodic function with period T. That is:

$$x(m) = x(m+T) \Rightarrow R(k) = R(k+T) \tag{8}$$

We can determine whether a speech signal is a voiceless or voiced consonant based on this property of the autocorrelation function. Furthermore, we can calculate the fundamental frequency period from a voiced speech signal. The autocorrelation fundamental frequency detection algorithm (ACF) precisely utilizes this property of the autocorrelation function R(k) to extract the fundamental frequency of the speech signal.

Above the threshold, multiple peaks appear. The distance between any two adjacent peaks

represents the fundamental period of the audio signal during that time interval, thereby enabling the calculation of the fundamental frequency for that period. This process constitutes the fundamental principle of using the Auto-Correlation Fundamental Frequency Detection Algorithm (ACF) to determine the fundamental frequency. The fundamental frequency extraction flowchart is shown in Figure 2.

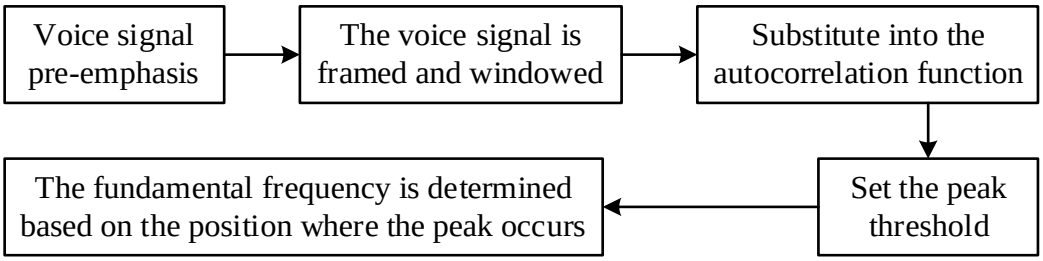


Figure 2: Flowchart of fundamental frequency extraction

3.2 Extraction of MFCCs

In singing competitions, contestants are required to articulate clearly and enunciate distinctly. There must be no errors in the expression (performance) of lyrics, meaning no mispronounced words. Since Mel Frequency Cepstral Coefficients (MFCC) accurately simulate human auditory characteristics and exhibit high noise resistance, MFCC is employed in this paper as a feature parameter for assessing singers' mastery of vocal techniques in both the singing technique timbre feature extraction method and the singing scoring method. For these reasons, MFCC is adopted in this paper as the feature parameter for evaluating singers' command of vocal techniques within the singing scoring methodology. Following the MFCC extraction procedure, the raw audio signal undergoes preprocessing. A Fast Fourier Transform (FFT) is then applied to obtain the preprocessed signal's time-domain waveform. This waveform is subsequently passed through a set of MEL filters to derive the Mel-frequency cepstral coefficients.

This paper employs MFCC for processing. MFCC parameters are extracted in the frequency domain using the Mel scale, accurately describing the nonlinear frequency response characteristics of the human ear. Their relationship with frequency can be approximated by the following equation:

$$\mathrm{Mel}(f) = 1125\ln(1+\frac{f}{700}) \tag{9}$$

In the formula, f represents frequency in Hz. The figure below illustrates the relationship between Mel frequency and linear frequency:

To effectively distinguish between these two distinct frequencies, the frequency signal must be confined to a specific bandpass range. This bandpass is termed the critical band, and its corresponding bandwidth is the critical bandwidth. The mathematical formula for calculation is as follows:

$$BW_C = 25 + 75\times[1+1.4\times(\frac{f_c}{1000})^2]^{0.69} \tag{10}$$

In the above equation, f_c represents the center frequency, measured in Hz.

From the formula, it can be seen that the critical bandwidth exhibits a growth trend

consistent with the perceived frequency. Specifically, when the signal frequency is below 1 kHz, the critical bandwidth increases approximately linearly [24]. When the signal frequency exceeds 1 kHz, the critical bandwidth follows an approximate logarithmic growth pattern. Based on this principle of critical bandwidth, a triangular filter bank can approximate an equivalent critical bandwidth filter bank. This triangular filter bank is known as the Mel filter bank, where the frequency corresponding to each vertex of the triangular filter bank represents the center frequency of that triangular filter. It is evident that the upper and lower limits of each triangular filter's frequency bandwidth fall precisely at the center frequencies of the two adjacent filters preceding and following it. Consequently, overlapping transition frequency bands exist between each pair of triangular filters, and the sum of their frequency responses equals 1.

$$H_m(k)=\begin{cases} 0 & k<f_{m-1}\ or\ k>f_{m-1} \\ \dfrac{k-f_{m-1}}{f_m-f_{m-1}} & f_{m-1}<k<f_m \quad m=0,1,2,\cdots .M-1 \\ \dfrac{f_{m+1}-k}{f_{m+1}-f_m} & f_m<k<f_{m+1} \end{cases} \tag{11}$$

In the above equation, f_{m-1}, f_m and f_{m+1} represent the upper cutoff frequency, center frequency, and lower cutoff frequency of the triangular filter, respectively. M denotes the number of triangular filters, i.e., the order of the Mel filter bank.

Based on the fundamental principles of MFCC parameters outlined above, the basic workflow for extracting MFCC feature parameters is as follows:

The specific computational steps are as follows:

(1) Perform preprocessing on the original music signal, including sampling and quantization, pre-emphasis, and windowing for frame segmentation. Assume the sequence of each obtained music signal frame is $s(n)$.

(2) Perform a Fast Fourier Transform (FFT) on the resulting sequence $s(n)$ to transform the music signal from the time domain to the frequency domain, yielding its spectrum. Let the spectrum of the transformed music signal be denoted as:

$$X_a(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi k/N},0\le k\le N \tag{12}$$

(1) The input speech signal in Equation $x(n)$, where N denotes the number of points in the Fourier transform.

(3) Calculate the energy spectrum, i.e., the square of the spectral amplitude.

(4) Pass the energy spectrum through a set of Mel filter banks $H_m(k)$. This paper adopts M=24.

(5) Perform logarithmic operations to compute the logarithmic energy output for each filter bank as:

$$S(m)=\ln\left(\sum_{k=0}^{N-1}|X_a(k)|^2H_m(k)\right),0\le m\le M \tag{13}$$

In the formula, M denotes the number of filters in the filter bank.

(6) To obtain MFCC parameters, perform a Discrete Cosine Transform (DCT) on S(m).

$$C(n)=\sum_{m=0}^{N-1}S(\mathrm{m})\cos\left(\frac{\pi n(\mathrm{m}\text{-}0.5)}{M}\right),0\mathrm{n}<\mathrm{M} \tag{14}$$

The order of MFCC coefficients is typically set between 12 and 16; this paper employs 16-order cepstral coefficients.

The feature vector set obtained through feature extraction must undergo normalization before use. Normalization ensures that each dimension possesses the same mean vector and variance.

For a piece of music, its acoustic signal is not static; it exhibits typical local transient characteristics. Since the Fourier transform does not effectively capture spectral information at specific times and frequencies, the short-time Fourier transform can be employed when studying music signals.

Performing a Fourier transform (discrete-time Fourier transform, DTFT) on the nth frame of audio signal $x_n(m)$ yields the short-time Fourier transform, defined as follows:

$$X_n(e^{j\omega})=\sum_{m=0}^{N-1}X_n(m)e^{-j\omega m} \tag{15}$$

From the above equation, it can be seen that the essence of the short-time Fourier transform is to apply a window function to the standard Fourier transform. This window function moves across the entire signal during the transformation process. It is precisely the addition of this window function that endows the original Fourier transform with the capability for local analysis.

The discrete short-time Fourier transform (DFT) is essentially a sampling of $X_n(e^{i\omega})$ in the frequency domain. In digital processing of audio signals, the discrete Fourier transform $X_n(k)$ of $x_n(m)$ is used to replace $X_n(e^{i\omega})$, and the transformation from $x_n(m)$ to $X_n(k)$ can be efficiently accomplished using the fast Fourier transform (FFT) algorithm.

3.3 Extracting Sound Intensity

We know that sound intensity represents the energy of a sound signal, determined by the amount of sound energy or acoustic pressure generated when an object vibrates. The greater the amplitude of the sound vibration, the greater the sound intensity, and the greater the loudness perceived subjectively by the human ear.

The sound intensity curve is defined as:

$$Vi=\frac{1}{M}\sum_{m=1}^{M}|S(m)| \tag{16}$$

Here, Vi denotes the average sound intensity of the entire audio file, $|S(m)|$ represents the signal amplitude at the mth sample point, and M is the number of sample points.

3.4 Feature Parameter Matching Method

Assuming the reference template is: R(1, 2, ..., M), and the target template is: T(1, 2, ..., N),

The DTW basic algorithm is as follows:

Similarity is represented by the distance D[T, R]. Let n and m denote arbitrary frame indices selected from T and R, respectively. D[T(n), R(m)] denotes the distance between these two frames. The DTW algorithm typically employs Euclidean distance, where a smaller distance indicates higher similarity.

If N=M, direct calculation is possible; otherwise, alignment of T(n) and R(m) must be considered. The objective of the DTW algorithm is to find an optimal time alignment function that nonlinearly maps the time axis n of the input signal template onto the time axis m of the reference template, minimizing the total cumulative distortion between them.

Here, we define the time alignment function as:

$$C = \{c(1), c(2), \cdots, c(i), c(I)\} \tag{17}$$

Here, I denotes the length of the matching path, i.e., the number of final matching points. c(i) = (R(i), T(i)) represents the matching pair formed at the i-th matching point, consisting of the R(i)-th feature parameter vector from the standard reference template and the T(i)-th feature parameter vector from the target sound signal template.

In the above formula, the time regularization function C should satisfy the following preliminary constraints:

Monotonicity:

$$R(i-1) \le R(i) \quad T(i-1) \le T(i) \tag{18}$$

Start-End Consistency:

In DTW matching calculations, it is generally required that:

$$R(1) = T(1) = 1 \quad R(I) = M \quad T(I) = N \tag{19}$$

Matching Continuity:

Generally, no matching points should be omitted, that is:

$$R(i) - R(i-1) \le 1 \quad T(i) - T(i-1) \le 1 \tag{20}$$

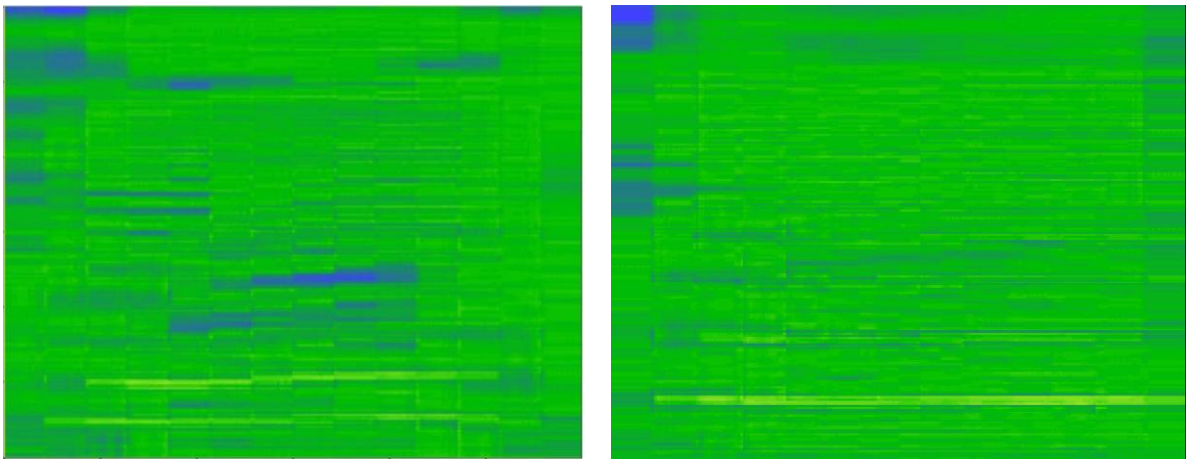
The final step involves matching the standard template with the target template. Here, we define the distance $d(R_{R(i),}T_{T(i)})$ between the standard template and the target template as the local matching error, also known as the local matching distortion value. The DTW algorithm achieves the minimum error value for the entire global optimization problem by solving each local optimization problem, that is:

$$D = \min_{c} \frac{\sum_{i=1}^{l} \left[d\left(R_{R(i),}T_{T(i)}\right) * W_i \right]}{\sum_{i=1}^{l} W_i} \tag{21}$$

4 Extraction of Timbre Characteristics in Vocal Techniques for Music Education

4.1 Spectrum Feature Extraction

This paper selected the Mel-spectrum feature map as the classification feature for distinguishing between high and low voices in men and women. The high-pitched Mel-spectrum diagram is shown in Figure 3 (Figure a represents male, Figure b represents female). In the diagram, the horizontal axis represents time, the vertical axis represents frequency, the brightest line at the bottom indicates the fundamental frequency, while the higher bright lines represent overtones. The color of the image indicates loudness information, with darker colors representing greater loudness. Comparing the high-frequency regions of male and female spectrograms reveals distinct patterns: even when derived from similarly pitched singing signals, female voices typically exhibit higher fundamental frequencies than males. Furthermore, female voices display richer high-frequency overtones with wider intervals between them, whereas male voices feature denser high-frequency overtones. These comparative differences provide a basis for classifying male and female voices.

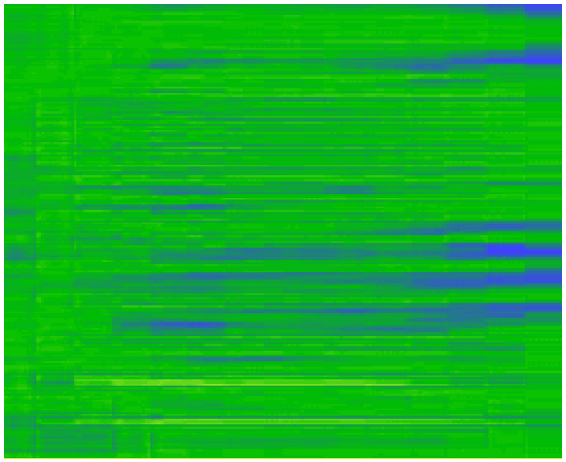


(a) Male (b) Female

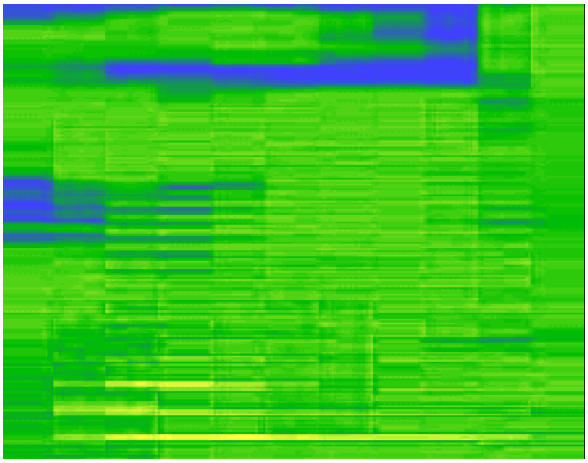
Figure 3: Treble spectrum

For the classification of high and low voices, the Mel-spectrograms of male voices are shown in Figure 4 (Figures a–c represent high, medium, and low voices, respectively). Taking male voices as an example, comparing the Mel-spectrum characteristics of tenor, baritone, and bass voices reveals distinct fundamental frequency ranges. Tenors exhibit the highest fundamental frequency and generally the greatest loudness, baritones have a lower fundamental frequency with reduced loudness compared to tenors, while basses possess the lowest fundamental frequency. Regarding the high-frequency harmonic components, the tenor exhibits the richest and densest high-frequency harmonics. The baritone's high-frequency harmonics appear relatively sparse, while the bass's harmonics are predominantly concentrated in the lower-middle portion of the image, showing the least high-frequency harmonics—even absent at the highest point of the image. By comparing treble, baritone, and bass voices, the distinct differences in their Mel-spectrogram features become evident, providing a solid foundation for pitch classification tasks across varying frequencies. The above analysis demonstrates that Mel-spectrograms exhibit distinct subtle variations across different genders and pitches. This confirms the feasibility of using Mel-spectrograms as input features for the extraction model to

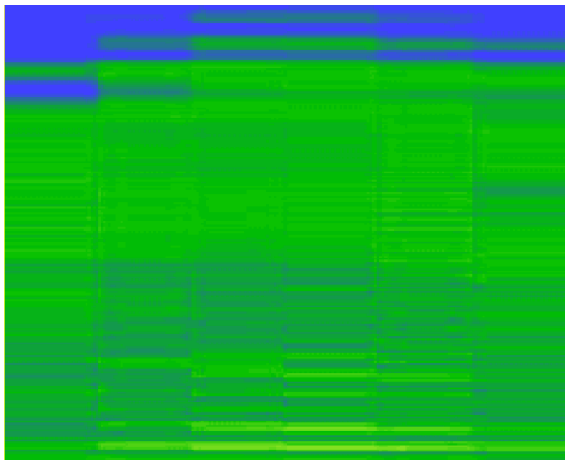
construct the musical timbre feature extraction model proposed in this paper.



(a) High note



(b) Alto



(c) Bass

Figure 4: Men's spectrum

4.2 A Cappella - Comparative Analysis of Initial Consonant Duration in Reading

The duration of initial consonants is short, accounting for a small proportion of the total syllable duration. When speech rate changes, the duration of initial consonants also changes. In singing, note durations vary according to musical notation requirements, with a greater range of variation. The range of variation in initial consonant duration and the definition of initial consonant boundaries are crucial for accurately analyzing note durations. Compared to normal-speed reading, initial consonant duration slightly extends when singing at varying speeds, though this extension is far less pronounced than for finals. Table 1 shows the duration relationship for different initial consonants in the reading-a cappella corpus during normal-speed reading and a cappella singing. Statistical analysis indicates that the degree of initial consonant extension correlates with the type of initial consonant.

Table 1: Reading aloud - Comparison of the duration of initial consonants

Initial consonant		Read the length (ms)	Singing length (ms)	Stretching ratio
Nonaspirated	b	22.37±6.55	25.7±4.85	1.1
	d	24.41±6.8	24.39±12.96	1.07
	g	39.43±8.97	43.55±21.1	1.07
Gas plug	p	47.97±5.77	53.77±15.91	1.06
	t	88.27±4.77	97.42±27.79	1.1
	k	83.74±19.47	94.27±15.99	1.09
Clear the fricative	s	115.19±2.58	172.22±0.85	1.49
	sh	85.18±6.5	124.17±42.68	1.47
	x	93.01±20.3	135.36±55.87	1.49
	f	77.88±21.07	114.58±12.95	1.48
	h	64.58±19.3	97.17±45.96	1.46
Non-Aspirated	z	77.21±11.01	115.67±30.08	1.52
	zh	52.67±3.58	79.45±17.13	1.49
	j	71.65±14.54	107.46±25.93	1.49
Scavenge	c	93.32±21.14	144.63±6.03	1.53
	ch	73.78±10.07	109.91±13.88	1.47
	q	71.36±15.22	107.5±19.05	1.49
Nasal tone	n	48.63±11.61	79±30.88	1.65
	m	38.86±15.45	61.16±27.02	1.59
Marginal voice	l	55.34±46.7	67.81±49.07	1.18
Retroflex sound	r	51.09±5.91	54.63±1.71	1.02

4.3 Analysis of Vowel Pitch in Singing Voices

(1) Analysis of Vowel Pitch Variation

In speech, syllable pitch is influenced by tone, intonation, rhythm, and stress. In singing, pitch is determined by the melody specified in the musical score. Maintaining consistency between the musical melody and the tonal pitch of lyrics is a consideration composers face when writing music. However, few compositions achieve perfect alignment between the two, and conflicts may even arise. The influence of tonal pitch on vowel pitch in recitation and a cappella singing is shown in Table 2. It can be seen that in recitation, the pitch range for first-tone syllables is approximately 1.534 semitones, fourth-tone syllables span 4.34 semitones, while rising-tone and departing-tone syllables reach as high as 8.281 and 8.42 semitones respectively. In a cappella singing, however, the pitch variation per syllable is only 1.672 semitones, roughly equivalent to the first-tone syllables in recitation.

Table 2: The influence of pitch on pitch in reading aloud and a cappella singing

	Yin Ping		Yang Ping		Shangsheng		castration	
	Read aloud	A cappella	Read aloud	A cappella	Read aloud	A cappella	Read aloud	A cappella
Standard deviation	0.029	0.034	0.087	0.029	0.154	0.034	0.173	0.041
Relative maximum	1.045	1.055	1.185	1.047	1.289	1.035	1.262	1.065
Relative minimum	0.933	0.938	0.926	0.911	0.761	0.947	0.754	0.948
Interval difference (semitone	1.534	1.587	4.34	1.931	8.281	1.603	8.42	1.672

(2) Vibrato Pitch Duration Analysis

In singing, vibrato is produced by oral airflow vibrations inducing vocal cord oscillations, resulting in periodic pitch and energy fluctuations. Proper application of vibrato enhances a song's artistic expressiveness. Vibrato may occur solely during the vowel portion of a note and extend until the syllable concludes. In the sight-singing corpus used herein, vibrato annotation was entirely manual. Annotators identified vibrato onset and offset points by observing pitch contours and cross-referencing audio recordings. Vibrato was only annotated for notes exceeding 500ms in absolute duration. The ratio of vibrato instances to total notes is presented in Table 3. The table indicates that the occurrence rate of trills increases with the absolute duration of notes. When note duration exceeds 1 second, 76.47% of notes exhibit trills.

Table 3: The comparison between the number of trill sounds and the total number

	0.6	0.8	1	1.2	1.4	1.6	1.8	>1.9
Notes without vibrato	109	17	9	9	7	1	0	6
Notes containing trills	44	19	25	19	12	21	5	22

4.4 Error Analysis

During spectral analysis, the following primary causes were identified for reduced accuracy:

(1) Low volume resulting in weak spectral energy, leading to missed detections

During audio recording, sudden volume drops may occur due to factors like distance between the speaker and microphone. Low volume causes missed detection samples, as shown in Figure 5. The initial sound at the starting point exhibits low energy in the spectrum.

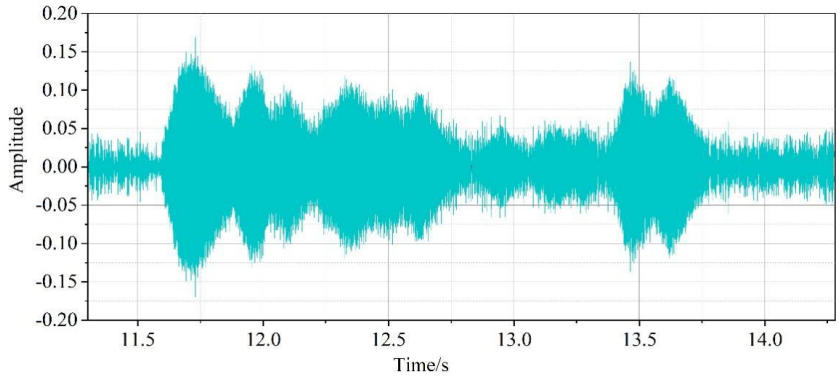


Figure 5: Low volume results in leakage samples

(2) Consecutive notes of the same pitch with similar energy levels may cause detection failures.

Sheet music often contains sequences of consecutive notes at identical pitches. As shown in Figure 6, samples of notes with the same pitch and similar energy levels exhibit indistinguishable starting points in the frequency spectrum. In the time domain, there is also no clear boundary between the two notes.

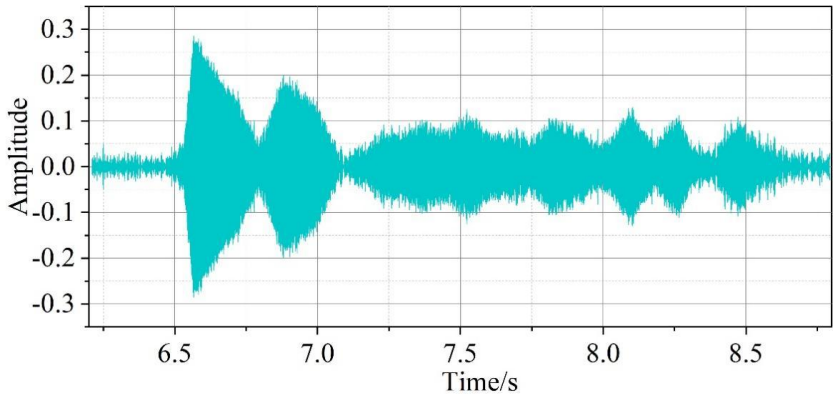


Figure 6: A sample of a note similar to that of a similar energy

(3) Excessive noise energy leads to false detections.

As shown in Figure 7, noise causes a false detection point in the spectrum plot. Correspondingly, in the time domain, a distinct noise spike is visible at that instant, which is caused by the metronome.

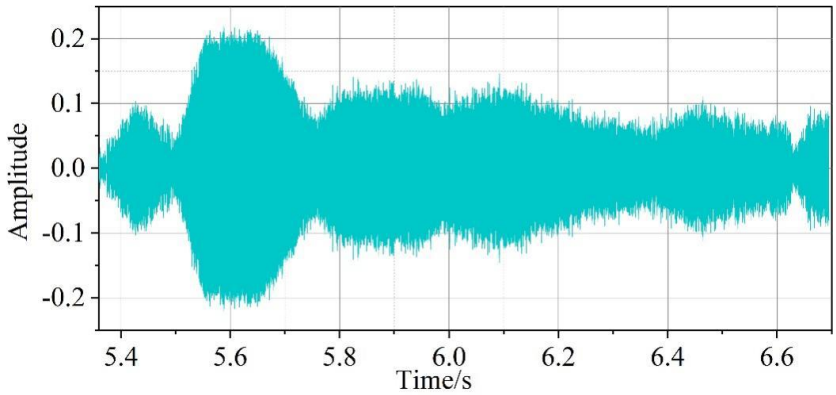


Figure 7: Noise leads to multiple sample detections

(4) Unstable breath control during singing causes inconsistent energy levels, leading to multiple detections.

As shown in Figure 8, samples with unstable breath control exhibit multiple detections. It is evident that when the singer produces this note, significant changes occur in the time-domain envelope. These changes manifest as fluctuations in energy levels in the frequency domain, resulting in multiple detections.

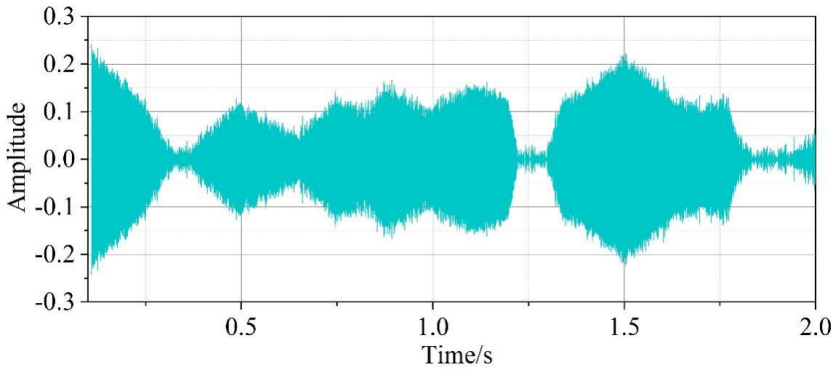


Figure 8: Unstable breath control leads to multiple samples being tested

5 Singing Assessment Strategies

5.1 Scoring Formula

The fundamental objective of the vocal scoring method studied in this paper is to identify objective musical skill metrics that measure a singer's mastery of lyrics, melody, and pitch during song performance, along with corresponding acoustic signal feature parameters. By comparing the feature parameter vectors of the singer's vocal samples against those of standard reference samples, the differences between them are calculated. Finally, these differences are synthesized into a scoring formula to calculate the singer's performance level score.

The original artist's vocal recording is used as the reference standard because any evaluation requires a benchmark. For most non-professional singers in real life, if their rendition of a song closely matches the original artist's performance, we can consider their singing proficiency high—essentially reaching professional standards. Conversely, if their rendition differs significantly from the original artist's, it clearly indicates a lower singing proficiency.

Some songs have been performed by multiple professional singers. To standardize audio files as reference samples, the pitch file of that song can serve as the standard vocal benchmark. By adjusting the weighting of metrics in the scoring formula below—such as reducing the weight of song content while increasing the weight of pitch accuracy—we can effectively reduce reliance on the original singer's voice and place greater emphasis on the singer's mastery of pitch and rhythm. Combining our factor evaluation approach, we derive the following singing scoring formula:

$$result = \frac{Vi\_B}{Vi\_A} * \left(\frac{k1 * d1}{1 + D(mfcc)} + \frac{k2 * d2}{1 + D(pitch)} \right) * 100 \tag{22}$$

Among these, k1 and k2 represent the weights assigned to lyrics and melody when a singer performs a specific song, with the condition that k1 + k2 = 1. d1 and d2 denote the difficulty coefficients for lyrics and melody respectively in song performance, and we stipulate that both d1 and d2 are greater than 1. Generally, songs with faster rhythms present greater difficulty in lyrics performance, while songs with higher overall pitch present greater difficulty in pitch performance. The weight and difficulty coefficient settings can be adjusted according to different requirements or scoring priorities.

In this experiment, we set kl=k2=0.5 and dl=d2=1. Thus, the above formula simplifies to:

$$result = \frac{Vi\_B}{Vi\_A} * \left(\frac{50}{1 + D(mfcc)} + \frac{50}{1 + D(pitch)} \right) \tag{23}$$

From the above formula, we can deduce:

When $D(mfcc) = D(pitch) = 0$, meaning the singer's lyrics and pitch throughout the entire song match the original singer perfectly, the maximum score is: $\frac{100 * Vi\_B}{Vi\_A}$

When $D(mfcc) = D(pitch) = \infty$, meaning the singer's lyrics and pitch throughout the entire song differ significantly from the original singer, the score will be very low, even approaching 0.

5.2 Vocal Performance Evaluation

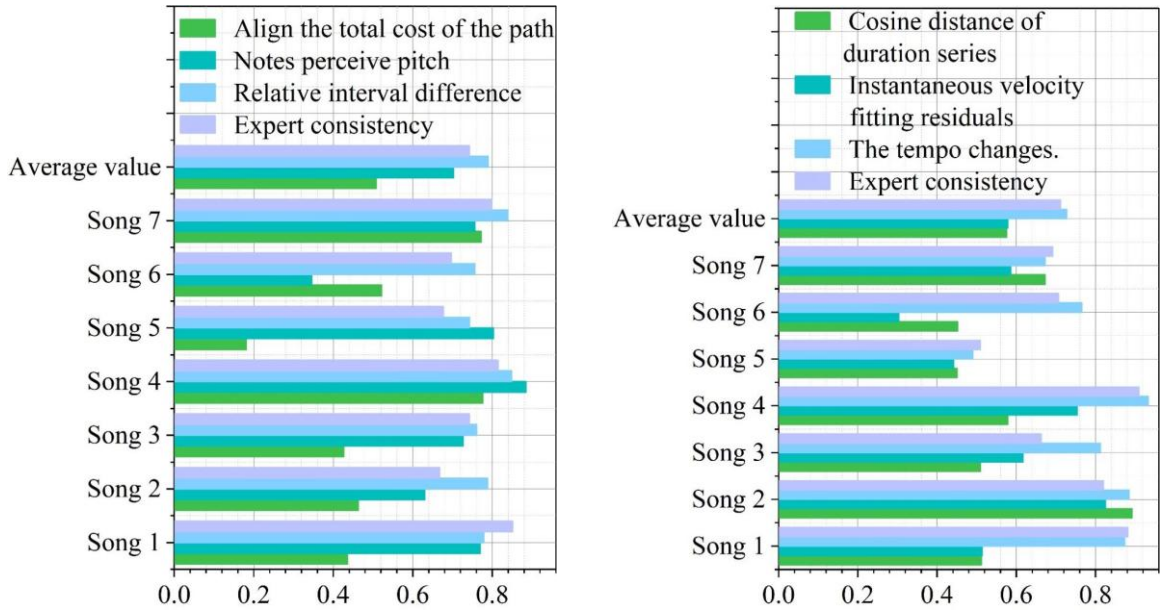
This section designs an evaluation experiment to examine the consistency between different assessment methods and expert evaluations regarding pitch and rhythm metrics. The experimental dataset comprises seven popular songs, each containing 25 to 55 musical notes. When performed at moderate tempo, the average duration is approximately 22 seconds. Six singers (three male, three female) participated in recording the experimental samples. All singers possessed basic professional vocal training, enabling them to maintain relatively accurate pitch and rhythm. However, occasional errors such as misreading sheet music, going off-key, or losing the beat occurred. During recording, singers performed the given musical notation by name. The system did not provide a reference pitch; singers independently chose either solfège or fixed-do solfège, with permission to transpose based on their vocal range. The sight-singing process was conducted without instrumental accompaniment or metronome use, requiring singers to independently control tempo and rhythmic timing. Singers were granted artistic freedom in musical expression, including the use of vibrato, legato, and other techniques. Each singer completed sight-singing for 7 songs, resulting in a total of $7 \times 6 = 42$ sight-singing audio recordings. Fifteen experts participated in the subjective evaluation scoring. All 15 experts received rigorous musical and aural training, possessing strong backgrounds in vocal, instrumental, or conducting disciplines. After listening to each vocal performance, experts scored both pitch accuracy and rhythmic precision on a granular scale (1–10 points). Due to significant variation in scoring scales among experts, this experiment employed the commonly used ranking method to evaluate sight-singing samples. Based on the experts' scores, each expert provided a ranking sequence for the six versions of the same song. The position (rank) of each sight-singing sample within the sequence was recorded, disregarding the specific scores. Subsequently, the order assigned by all experts for the same sight-singing sample was averaged to determine the average rank for each sample. Finally, the average ranks of different samples for the same song were reordered to generate a final sequential ranking scale. To minimize significant rank discrepancies among samples of similar quality, identical rankings were permitted. When multiple samples tied, the final rank was the average of their total-order rankings. For each sight-singing sample, automated annotation methods detect note onset positions and extract pitches. Parameters required for three objective singing evaluation criteria are calculated against the reference score, ultimately outputting Pitch Deviation (ID) and Rhythm Deviation (RD). Lower deviation values indicate closer alignment with the reference score, resulting in higher sight-singing sample scores. Each song's six versions are ranked based on system-assigned scores for both pitch and rhythm evaluation criteria. The final assessment focuses solely on the relative order among different versions of the same song. Since deviation values are not quantified, no ties occur between sight-singing samples. The entire evaluation process operates autonomously without expert guidance or manual intervention. The correlation between the singing assessment method and expert subjective evaluations is shown in Table 4. It is evident that the objective evaluation criteria based on note-level features demonstrate superior performance in measuring pitch accuracy. Their correlation coefficient with experts is 0.6984, approaching the average inter-expert agreement and significantly surpassing evaluation standards based on frame-level acoustic features. This indicates that human perception of music operates at the note level, discerning pitch through the overall perception of internally stable pitch. Regarding rhythm, both the note-level evaluation criteria and current methods exhibit suboptimal performance. In contrast, the correlation coefficients between the evaluation results based on the proposed method and expert assessments reached 0.7826 for pitch and 0.7687 for rhythm. These values surpass the average correlation coefficient among the ten experts. This demonstrates that the beat and pitch information extracted using the proposed method can effectively detect deviations between singing samples and reference scores, yielding reasonable

scores.

Table 4: The correlation between the method and the expert subjective evaluation

Song	Evaluation criteria	Evaluation standard based on acoustic characteristics	Based on the characteristics of the notes	Evaluation methods	The consistency between expert evaluation
Song 1	Pitch	0.4398	0.7649	0.7639	0.8455
	Rhythm	0.4931	0.5149	0.8687	0.8684
Song 2	Pitch	0.4502	0.6254	0.7812	0.6434
	Rhythm	0.8525	0.8049	0.8679	0.801
Song 3	Pitch	0.4261	0.7353	0.7698	0.7415
	Rhythm	0.4932	0.6182	0.792	0.6565
Song 4	Pitch	0.7908	0.8849	0.8387	0.8045
	Rhythm	0.5771	0.7544	0.9184	0.9189
Song 5	Pitch	0.1863	0.8002	0.7341	0.6713
	Rhythm	0.4411	0.4368	0.4912	0.5123
Song 6	Pitch	0.507	0.3474	0.7538	0.681
	Rhythm	0.4522	0.3009	0.7699	0.6857
Song 7	Pitch	0.7727	0.7304	0.8366	0.7839
	Rhythm	0.6598	0.5838	0.673	0.6858
Mean of song	Pitch	0.5104	0.6984	0.7826	0.7387
	Rhythm	0.5670	0.5734	0.7687	0.7327
Overall average value		0.5387	0.6359	0.7757	0.7357

When sight-singing samples exhibit significant differences in quality, experts can readily provide accurate rankings. Conversely, when sample variations are subtle or difficult to compare, expert judgments become susceptible to subjective perceptions and external noise. The effectiveness of this vocal assessment method and inter-expert consistency are illustrated in Figure 9 (Figure a shows pitch accuracy metrics, Figure b shows rhythm accuracy metrics). Statistical analysis reveals that within the pitch criteria, the correlation coefficients between the effectiveness of alignment path cost, perceived note pitch, and relative note pitch metrics and the distinguishability of sight-singing samples are 0.6043, 0.4145, and 0.6254, respectively. For rhythm-related criteria, the correlation coefficients between the effectiveness of duration sequence cosine distance, instantaneous velocity fitting curve residual, and beat velocity variation metrics and the distinguishability of sight-singing samples were 0.3922, 0.5121, and 0.9105, respectively. It can be observed that, regardless of whether it is pitch accuracy or rhythm accuracy, the effectiveness of singing scores in both pitch and rhythm aspects shows a strong positive correlation with the distinctiveness of the sight-singing sample set itself.



(a) Pitch accuracy measurement (b) Rhythm accuracy measurement

Figure 9: The validity of this method is compared with the experts

6 Conclusion

With the growing emphasis on the music education industry, research into extracting vocal timbre characteristics from singing techniques in music education is gaining increasing attention. This paper proposes a method for extracting vocal timbre characteristics from singing techniques in music education, alongside a singing evaluation approach. The conclusions drawn are as follows:

(1) Comparative analysis of a cappella singing versus spoken recitation reveals that alterations in syllable duration primarily stem from vowel-based elongation. Vowel pitch remains relatively stable, determined by musical notation, with Chinese tonal patterns exerting minimal influence on intrasyllabic pitch variation.

(2) Correlation analysis between objective evaluation criteria and expert subjective assessments indicates that the proposed method demonstrates strong performance in measuring pitch accuracy. Its correlation coefficient with expert evaluations (0.6984) approaches the average inter-expert agreement and significantly exceeds evaluation standards based on frame-level acoustic features. This confirms that the proposed method effectively reflects singers' proficiency in song pitch and lyrics, with evaluation results aligning with human subjective perception.

About the Author

Wenjing Wu: Born in December 1982, female, Han ethnicity, Hebi City, Henan Province, China, PhD, Associate Professor, Music Education Management

Jacquline Tham: Personal Profile: Born in February 1972, female, Selangor, Malaysia, PhD, Professor, Education Management

References

[1] Zhu, X., Beauregard, G. T., & Wyse, L. L. (2007). Real-time signal estimation from modified short-time Fourier transform magnitude spectra. IEEE Transactions on audio, speech, and language processing, 15(5), 1645-1653.

[2] Nieto, O., Mysore, G. J., Wang, C. I., Smith, J. B., Schlüter, J., Grill, T., & McFee, B. (2020). Audio-based music structure analysis: Current trends, open challenges, and applications. Transactions of the International Society for Music Information Retrieval, 3(1).

[3] Desai, D., & Mehendale, N. (2022). A review on sound source localization systems. Archives of computational methods in engineering, 29(7), 4631-4642.

[4] Gupta, V., Mittal, M., Mittal, V., & Saxena, N. K. (2021). A critical review of feature extraction techniques for ECG signal analysis. Journal of The Institution of Engineers (India): Series B, 102(5), 1049-1060.

[5] Van Elferen, I. (2021). The Vibrant Aesthetics of Tone Color. The Oxford Handbook of Timbre, 69.

[6] Szymański, F., Łukasik, E., & Chudy, M. (2024). Timbra: An online tool for feature extraction, comparative analysis and visualization of timbre. International Journal of Electronics and Telecommunications, 349-354.

[7] Su, Y. (2024, June). Similarity of Musical Timbres Using Fourier Transform. In 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA) (pp. 450-453). IEEE.

[8] Jacobsen, S., & Siedenburg, K. (2024). Exploring the relation between fundamental frequency and spectral envelope in the perception of musical instrument sounds. Acta Acustica, 8, 48.

[9] Gonzalez, Y., & Prati, R. C. (2022). Acoustic descriptors for characterization of musical timbre using the fast Fourier transform. Electronics, 11(9), 1405.

[10] McAdams, S., Thoret, E., Wang, G., & Montrey, M. (2023). Timbral cues for learning to generalize musical instrument identity across pitch register. The Journal of the Acoustical Society of America, 153(2), 797-811.

[11] Chen, K. Y., Ding, J. J., & Lee, Y. K. (2024, June). Mixed Music Instrument Classification Based on Instantaneous Frequency Analysis and Time-Variant Spectrum Information. In Proceedings of the 2024 8th International Conference on Graphics and Signal Processing (pp. 30-35).

[12] Sun, Y. (2023). Timbre-Based Portable Musical Instrument Recognition Using LVQ Learning Algorithm. Mobile Networks and Applications, 28(6), 2171-2181.

[13] Blaszke, M., & Kostek, B. (2022). Musical instrument identification using deep learning approach. Sensors, 22(8), 3033.

[14] Gonzalez, Y., & Prati, R. C. (2023). Similarity of musical timbres using fft-acoustic descriptor analysis and machine learning. Eng, 4(1), 555-568.

[15] Li, R. (2022). Intelligent analysis of music education singing skills based on music waveform feature extraction. Mobile Information Systems, 2022(1), 9747342.

[16] Huang, C. (2022). Vocal music teaching pharyngeal training method based on audio extraction by big data analysis. Wireless Communications and Mobile Computing, 2022(1), 4572904.

[17] Wei, C. (2021). A Study of piano timbre teaching in the context of artificial intelligence interaction. Computational Intelligence and Neuroscience, 2021(1), 4920250.

[18] Wei, Y., Gan, L., & Huang, X. (2022). A review of research on the neurocognition for timbre perception. Frontiers in Psychology, 13, 869475.

[19] Jiang, W., Liu, J., Li, Z., Zhu, J., Zhang, X., & Wang, S. (2019, June). Analysis and modeling of timbre perception features of chinese musical instruments. In 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS) (pp. 191-195). IEEE.

[20] Acquilino, A., Puranik, N., Fujinaga, I., & Scavone, G. P. (2023). A Dataset and Baseline for Automated Assessment of Timbre Quality in Trumpet Sound. In ISMIR (pp. 684-691).

[21] Wang, R., Qi, J., & Qiao, D. (2022, July). An online vocal music teaching timbre evaluation method based on feature comparison. In International Conference on E-Learning, E-Education, and Online Training (pp. 482-494). Cham: Springer Nature Switzerland.

[22] Di Stefano, N. (2023). Musical emotions and timbre: From expressiveness to atmospheres. Philosophia, 51(5), 2625-2637.

[23] Verma, T., Aker, S. C., & Marozeau, J. (2023). Effect of vibrotactile stimulation on auditory timbre perception for normal-hearing listeners and cochlear-implant users. Trends in Hearing, 27, 23312165221138390.

[24] J. Jayanthi & V. Upendran. (2025). Raga Recognition of Indian Classical Music using Meerkat Optimization Based MFCC and Fine Tuned BILSTM-XGBOOST. Circuits, Systems, and Signal Processing,44(7),1-29.