International Journal of Computer Applications in Technology

ISSN online: 1741-5047 - ISSN print: 0952-8091
https://www.inderscience.com/ijcat

An extraction method of pop music singing beats based on audio features

Zhuo Kong, Guofeng Liu

DOI: 10.1504/IJCAT.2026.10075713

Article History:

Received:	23 July 2025
Last revised:	13 November 2025
Accepted:	14 November 2025
Published online:	22 May 2026

An extraction method of pop music singing beats based on audio features

Zhuo Kong and Guofeng Liu*

Conservatory of Music,
Vocal Music Department,
Jiamusi University,
Jiamusi, Heilongjiang Province, China
Email: 455688777@qq.com
Email: 658696@163.com
*Corresponding author

Abstract: In the analysis process of popular music singing audio, factors such as environmental noise interference and complex instrument accompaniment seriously affect the accuracy of audio feature extraction, resulting in the performance of traditional music beat extraction methods being difficult to meet practical needs. Therefore, this study innovatively proposes a popular music singing beat extraction method based on multifeature fusion. Performing pre-processing operations such as discretisation, denoising and normalisation on the original singing audio signal effectively improves signal quality. Through joint time-frequency domain analysis, comprehensively extract the time-frequency characteristics of music signals. Adopting a feature fusion strategy, combined with beat cycle analysis and inter beat distance calculation, high-precision beat detection is achieved. Experimental data shows that the missed detection rate and false detection rate of this method are as low as 2.1% and 2.5%, respectively, significantly better than traditional methods, providing reliable technical support for pop music performance analysis.

Keywords: audio features; pop music; singing rhythm; intelligent extraction model.

Reference to this paper should be made as follows: Kong, Z. and Liu, G. (2026) 'An extraction method of pop music singing beats based on audio features', *Int. J. Computer Applications in Technology*, Vol. 78, No. 6, pp.1–10.

Biographical notes: Zhuo Kong received her Master's degree in Education from the University of Canberra in 2013 and PhD degree in Education in 2023. She is an Associate Professor in the Teaching and Research Office of Pop Music at the Conservatory of Jiamusi University, Heilongjiang Province. His main research interests include pop singing and teaching.

Guofeng Liu received PhD degree in Music Education (Philosophy). He is the Dean of The School of Music at Jiamusi University, Heilongjiang Province. His main research interests include vocal music singing and teaching.

1 Introduction

In the field of music, beat, as the core element of musical rhythm, constitutes the basic framework of music composition. The pattern of alternating strong and weak beats runs through the entire music performance process, endowing music with rich dynamics and rhythmic beauty. In popular music singing, rhythm not only serves as an important foundation for musical structure but also shows close correlation with vocal techniques, emotional expression and stage performance. With the rapid advancement of artificial intelligence technology, beat extraction technology has gradually become a pivotal component in music signal processing, music information retrieval, music generation and multimodal applications (Zhong et al., 2023). Currently, the accuracy and real-time capability of beat extraction algorithms directly influence the expressiveness and interactivity of musical works, demonstrating broad application potential in fields such as musical accompaniment, dance choreography and music education. However, existing technologies still face challenges in handling complex rhythmic variations, tempo-changing music and polyphonic compositions (Liu, 2024). Under these circumstances, in-depth research on beat extraction technology in popular music singing could not only help overcome the performance limitations of beat recognition algorithms and promote the integration and development of music technology with other disciplines, but also provide scientific support for music creation, performance and therapy, which holds significant importance for enriching music theory systems and enhancing the digitalisation level of the music industry.

In recent years, with the deep integration of artificial intelligence and music information processing technology, music beat extraction and its related applications have

become a research hotspot in the academic community. Many scholars have conducted research on the relationship between music rhythm and audio signal characteristics, and have achieved certain research results. Long et al. (2024) proposed an audio recognition method based on Self-Organising Map (SOM) and pulse Neural Network (SNN). The method compresses the spatiotemporal features of the audio signal through MFCC feature extraction and SOM sparse encoding, and optimises the SNN weights using STDP learning rules and excitation suppression dual supervision mechanism. Although this method has the ability to jointly extract spatiotemporal features at the theoretical level, its practicality is limited due to the high dependence of SNN's pulse issuance mechanism on the sparsity of audio signals, which significantly reduces the robustness of beat recognition in complex acoustic environments such as multi instrument mixed scenes. For example, Yang (2024) designed an artificial intelligence-based music rhythm recognition system, which extracts audio signal fingerprint features through complex cepstral and substring matching techniques, and compares them with a database for recognition. However, due to the high sensitivity of fingerprint features to local changes in audio signals, the stability of beat recognition in dynamic music scenes is insufficient, making it difficult to meet high-precision application requirements. Wang (2023) proposed a hybrid model that combines attention mechanism and Long-Short-Term Memory (LSTM) network, which dynamically generates performance actions by capturing the temporal dependencies of music beat sequences. In terms of technical implementation, the research team introduced attention mechanism and reconstructed a non-recursive RNN abstract framework based on the traditional image description generation model, further optimising the processing ability of LSTM network for long sequence data. However, when deployed in edge server architecture, the lack of dynamic adaptability in computing resource allocation strategy exacerbates the contradiction between data transmission delay and model inference efficiency, ultimately leading to a significant decrease in the synchronisation between action generation and music rhythm, limiting its application effectiveness in real-time interactive scenarios. Zhu (2022) focused on data-driven beat action matching error recognition and proposes a feature extraction and matching method based on statistical learning. This method models dance movement features as a set of positive and negative samples, calculates sample weights and absolute value features of beat signals, combines Gaussian filters to suppress noise interference, and finally uses Short-Time FourierT (STFT) to extract matching features and construct a Support Vector Machine (SVM) classification model. However, due to insufficient modelling of the temporal dynamics of beat signals in the feature extraction process, there is a deviation in the temporal correspondence between beat features and dance movements, and the final matching accuracy is difficult to meet the high-precision application requirements.

The feature extraction of popular music singing audio faces significant challenges from environmental noise interference and complex accompaniment, which directly affects the accuracy of beat detection. The existing methods have shortcomings in utilising time-frequency features and noise robustness, making it difficult to meet practical application requirements. To address this issue, our research aims to propose a novel audio feature analysis method that improves the robustness and accuracy of beat detection by optimising the pre-processing process and improving the feature extraction strategy. We intend to establish a complete framework for extracting singing beats, achieve high-precision beat detection in noisy environments and provide technical support for music information retrieval and intelligent accompaniment systems. The research technical route of this article is as follows:

(1) We have innovatively established a complete music signal processing flow, achieved multi-resolution analysis through discrete wavelet transform and effectively separated beat features from environmental noise by combining adaptive threshold denoising technology. While preserving the key features of the signal, a dynamically adjusted contraction coefficient is used for noise suppression, and device differences are eliminated through amplitude normalisation processing. The proposed approximate coefficient preservation and detail coefficient contraction strategies, specifically targeting the energy accumulation characteristics of beat signals, significantly improve signal quality. This method breaks through the limitations of traditional techniques in noise robustness and provides a high-quality pre-processing foundation for subsequent beat feature extraction. It achieves excellent denoising effects while preserving the time-frequency characteristics of the music signal.

(2) A multi-dimensional feature extraction system for music singing audio has been constructed, which comprehensively captures beat features through time-frequency domain joint analysis method. We propose an improved short-term energy calculation model in time-domain analysis, and optimise zero crossing rate detection by combining dynamic window function design; In terms of frequency domain analysis, innovative use of complex cepstral transform technology effectively separates excitation sources from channel response characteristics. By establishing a collaborative analysis framework of short-term power spectrum and cepstral features, accurate characterisation of the time-frequency characteristics of music signals has been achieved. This method breaks through the limitations of single domain analysis and provides richer and more robust feature representations for beat extraction in complex music environments. Experimental verification shows that this feature extraction scheme significantly improves the accuracy and stability of subsequent beat detection.

(3) By constructing a time-frequency characteristic map matrix and using time bending technology to handle speed differences, we innovatively combine the feature analysis of harmonic components and impulse components. We establish a binary saliency model based on

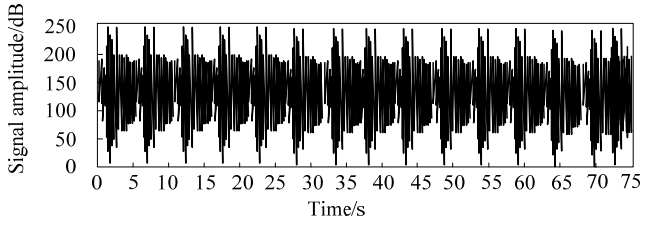
global periodic vectors and introduce a dynamic update mechanism in rhythm value estimation, integrating rhythm adaptive adjustment strategy into the design of interval distance function, and achieving optimal beat sequence search through dynamic programming. The method significantly improves the accuracy of beat detection in complex environments and achieves collaborative optimisation of feature analysis and beat tracking. The core innovation lies in the construction of feature map matrix, joint saliency modelling and adaptive beat distance measurement, providing new ideas for music beat extraction.

2 Pre-processing of audio features for pop music performance

2.1 Music singing audio signal processing

Pop music consists of interwoven sound elements with multiple frequencies, such as plucked guitar strings, drum beats or vocal pronunciations, each possessing unique frequency, amplitude, duration and spectral characteristics. When these elements combine through regular alternation of strength and periodicity, they form a harmonious musical structure (Jing, 2024; Fu, 2024). As the core element of temporal organisation in music, the beat fundamentally reflects variations in the strength and periodicity of different characteristic sounds. Its core features can be summarised as periodicity and continuity: periodicity manifests as the beat taking a fixed sequence of strong and weak sounds as its basic unit, while continuity manifests as the beat maintaining a fixed time interval, creating a repetitive stress pattern. Taking drum beats as an example, when the drummer strikes the drumhead, the vibration produces sound waves that exhibit regularity in loudness, intensity and time intervals (Han et al., 2024). Since beats typically coincide with various sound events, the total acoustic energy at beat points is significantly higher in localised time ranges compared to other periods (Guo and Wang, 2024). In encoded music signals, beat positions appear as convergence points of energy peaks. The waveform of the music signal is illustrated in Figure 1.

Figure 1 Waveform of music signal



As shown in Figure 1, the beat of popular music is hidden within these peaks. Based on this, complete the collection of popular music signals. Owing to the possible presence of environmental noise such as equipment background noise, electromagnetic interference and non-beat related frequency components such as instrument overtones and human voice harmonics in the initial signal, these factors can interfere with the effective extraction of beat features. Therefore, pre-processing of the original signal is necessary (Dong, 2022; Potempski et al., 2022). Assuming the music signal is $x(t)$ and the sampling frequency is f_s, the signal is first discretised to obtain a discretised signal $x[n]$, which can be expressed as:

$$x[n] = x(nT_s), n = 1, 2, 3, ... \tag{1}$$

In the formula, T_s is the sampling interval, which can be represented as $T_s = \frac{1}{f_s}$ (Li, 2024).

Then, based on wavelet analysis, the music signal is denoised. This time, discrete wavelet signals are used to perform multi-scale decomposition on the discretised signals, and the decomposition formula is as follows:

$$A_j[n] = \sum_k x[n] h_{k-2n} A_{j-1}[k] \tag{2}$$

$$D_j[n] = \sum_k x[n] g_{k-2n} A_{j-1}[k] \tag{3}$$

In the formula, $A_j[n]$ represents the approximation coefficient; $D_j[n]$ represents the coefficient of detail; h represents a low-pass filter; g represents a high pass filter; j represents the decomposition scale (Lan et al., 2024). After completing signal decomposition, threshold processing is applied to the detail coefficients of each scale to remove noise. The formula is as follows:

$$\bar{D}_{j,k}[n] = \begin{cases} \operatorname{sgn}(D_j[n]) \left(|D_j[n]| - \dfrac{\alpha |D_j[n]|}{\exp\left[m |D_j[n]| - t \right]^2} \right), \\ |D_j[n]| \geq t \\ \operatorname{sgn}(D_j[n]) \dfrac{(1-\alpha)|D_j[n]|}{\exp\left[|D_j[n]| - t \right]^2}, |D_j[n]| < t \end{cases} \tag{4}$$

In the formula, $\bar{D}_{j,k}[n]$ represents the result of $D_j[n]$ after threshold processing; t is the set threshold; $sgn(.)$ represents the sign function; α represents the coefficient of contraction; m represents the convergence speed adjustment coefficient (Zhao et al., 2023).

Performing wavelet inverse transform on the obtained approximate coefficients and the detail coefficients after threshold processing to achieve signal reconstruction, the results are as follows:

$$\dot{x}[n] = S \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{j,k} \bar{D}_{j,k}[n] \tag{5}$$

In the formula, S represents a constant coefficient; $C_{j,k}$ represents a discrete wavelet sequence.

Finally, the denoised signal is normalised to eliminate the gain differences between different recording devices, so that the signal has a uniform amplitude range (Yu et al., 2023), which can be expressed as:

$$X[n]=\frac{\dot{x}[n]-\min \dot{x}[n]}{\max \dot{x}[n]-\min \dot{x}[n]} \tag{6}$$

In the formula, $\max\dot{x}[n], \min\dot{x}[n]$ represents the maximum and minimum values of $\dot{x}[n]$.

Based on the above, complete the processing of music singing audio signals to lay the foundation for subsequent audio signal feature extraction and beat extraction.

2.2 Feature extraction of music singing audio

After completing the pre-processing of popular music singing audio signals, higher purity, less noise interference, and clearer spectral characteristics of audio signals can be obtained, thereby improving the accuracy and effectiveness of feature extraction. Next, we will carry out the design of feature extraction for music singing audio, as follows:

1) *Time domain characteristics*: The time-domain characteristics of music singing audio signals include short-term average energy, short-term average zero crossing rate and other indicators. This time, it is believed that the short-term average energy is a set of time series, which is obtained by squaring the amplitude of the sampling points of a frame of music signal and then passing through an impulse response to obtain the output from the filter of h (Yang, 2023). The formulas for extracting the short-term average energy of the music signal at time n are as follows:

$$E_n=\sum_{n=n-M+1}^{N} X^2[n]h(n-M) \tag{7}$$

In the formula, M represents the window length and h represents the impulse response.

The short-term average zero crossing rate is the ratio of the number of times the sampling point values of an audio signal change from positive to negative or from negative to positive within a short-term analysis window (i.e., the number of zero crossings) to the total number of sampling points within the window. It reflects the frequency characteristics of the signal in the time domain (Han et al., 2023), and its extraction formula is as follows:

$$Z_n=\frac{1}{2N}\left|\operatorname{sgn}[X(n)]-\operatorname{sgn}[X(n-1)]\right| * w(n) \tag{8}$$

In the formula, $w(n)$ represents the window function; $\operatorname{sgn}[.]$ represents the sign function, which is set as follows:

$$\operatorname{sgn}[X(n)]=\begin{cases}1, X(n)\geq 0\\ -1, X(n)<0\end{cases} \tag{9}$$

2) *Frequency domain characteristics*: The frequency domain characteristics of music singing audio signals include short-time power spectrum and cepstral. The so-called short-time power spectrum refers to the Fast Fourier Transform (FFT) of an audio signal within a very short time window (usually tens of milliseconds) to obtain the energy distribution of the signal on various frequency components within that time window. It reflects the power magnitude of the audio signal at different frequencies in a short period of time, similar to slicing a piece of music at a microscopic time scale and analysing the frequency energy situation in each 'slice' (Yang and Li, 2023). The extraction process is as follows:

$$P_n(w)=\sum_k R_n(k)e^{jwk} \tag{10}$$

In the formula, $R_n(k)$ represents the autocorrelation function of the audio signal, which can be expressed as:

$$R_n(k)=\sum_{m=n}^{n+N-k-1} x_w(m)x_w(m+k) \tag{11}$$

The so-called cepstral is a feature representation obtained by inverse Fourier transform of the logarithmic power spectrum of a signal, which can extract the excitation source information and channel response information of the signal (Wu et al., 2023). The extraction process is as follows:

When using the Z transform to represent the short-time Fourier transform, if there is a signal $X(n)$, the logarithm of the transformed signal Z is:

$$\bar{X}(z)=\ln\left[Z[X(n)]\right] \tag{12}$$

The inverse Z transformation can be written as:

$$\dot{x}(n)=Z^{-1}\left[\ln Z[X(n)]\right] \tag{13}$$

When taking $z=e^{jw}$, the above formula can correspond to the complex cepstral domain:

$$\dot{Z}(e^{jw})=\ln\left|X(e^{jw})\right|+j\arg\left[X(e^{jw})\right] \tag{14}$$

Namely:

$$\dot{x}(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\dot{X}(e^{jw})e^{jwn}dw \tag{15}$$

$\dot{x}(n)$ is the complex cepstral representation of $X(n)$. When taking the logarithm of the absolute value of $X(e^{jw})$, we have:

$$\dot{X}(e^{jw})=\ln\left|X(e^{jw})\right| \tag{16}$$

The inverse frequency spectrum $c(n)$ obtained from this is the real cepstral, abbreviated as cepstral, which can be expressed as:

$$R_n(k)=\sum_{m=n}^{n+N-k-1} x_w(m)x_w(m+k) \tag{17}$$

In summary, it can achieve feature extraction of music singing audio, laying the foundation for subsequent pop music singing rhythm extraction.

This study innovatively constructed a multi-level feature extraction topology structure for music singing audio, and established a signal decomposition framework based on discrete wavelet transform from the perspective of time-frequency dual domain collaborative analysis. At the level of feature extraction algorithms, a short-term energy analysis method with dynamic window length adjustment is proposed, combined with an improved zero crossing rate detection model, to achieve accurate capture of time-domain features; The creative application of complex cepstral transform technology in frequency domain analysis effectively enhances the feature representation ability through the spectral separation mechanism of excitation source and channel response. In the implementation process, multi-scale wavelet threshold denoising pre-processing is adopted, and a short-time power spectrum analysis module with time-varying window function is constructed. A feature fusion strategy of logarithmic spectral domain inverse transform is designed to form a complete computational link from the original signal to high-order features. This feature extraction system provides multi-dimensional discrimination basis for music beat analysis through deep coupling of time-frequency features.

3 Design of extracting rhythm for popular music singing

Completing the feature extraction of music singing audio can obtain a series of feature parameters that can characterise different characteristics of audio signals, providing rich data foundation and effective feature representation for subsequent music beat extraction. Next, by integrating the audio features of popular music singing, we will complete the design of extracting the rhythm of popular music singing, as follows:

Firstly, combining the time-frequency domain characteristics of the music signal, calculate the feature map matrix $TG^{v}(t,s)$ for each feature. The principle for determining the value of each element in the matrix is as follows: calculate the significance level in frame s under the condition that the current rhythm value is t, and take its maximum amplitude as the output of frame s. When the speed value of the music is low, the number of segments will correspondingly decrease. In order to maintain the same dimension between the feature map matrix TG^{v} at low speeds and the matrix at high speeds, it is necessary to use the Time Warping method to align the two when necessary. Secondly, with the above concept in mind, for each segment index s, the feature map matrices corresponding to the harmonic component and impulse component feature vectors are calculated as the chromaticity feature matrix TG^{ch} and the filtering energy feature matrix TG^{fl}. The global periodic vector of the entire music segment to be estimated is denoted as T_{gl}, which is represented by the product of TG^{ch} and TG^{fl} after their respective superposition processing. That is:

$$T_{gl}=\left(\sum_{S}TG^{fl}(t,s)\right)\left(\sum_{S}TG^{ch}(t,s)\right) \tag{18}$$

Perform significance analysis on T_{gl}, where the maximum value point is the corresponding global period. The analysis of rhythm values in this article is based on the theoretical foundation that the peak value in the global cycle vector is an integer multiple of a certain rhythm value. This rhythm value is only an intermediate variable used in the later division of beats, and is a scalar with physical meaning but no practical meaning. The calculation formula for the basic rhythm value T_0 corresponding to the global cycle is:

$$T_0=\arg\max_{t}\left\{\sum_{k}T_{gl}(kt)\right\} \tag{19}$$

To make T_{gl} reach its peak at both the target rhythm value and an integer multiple of T_0. To achieve this goal, establish a binary model $\{T_s, T_f\}$, where T_s is the fundamental variable and T_f is a multiple of T_s. The formula for calculating the joint significance J_s of the two variables is as follows:

$$J_s\{T_s,T_f\}=\left[T_{gl}(T_s)+T_{gl}(T_f)\right]\cdot\sum e^{-\left(T_f/T_s-i\right)^2/(p_i)^2} \tag{20}$$

Obviously, as T_s and T_f increase and the latter approaches a multiple of the former, T_s increases, and the final rhythm value T is T_s, which can maximise J_s and this must also be a multiple of T_0, that is:

$$T=\arg\max_{jT_0}\{J_s\{iT_0,kT_0\}\} \tag{21}$$

On the basis of obtaining the beat cycle, perform beat point tracking to obtain the position of the beat point in the audio signal. $r_{ch,T}^{k}$ and $r_{fl,T}^{m}$ represent all peaks of feature sequences x_{ch}^{k} and x_{fl}^{m} with respect to beat velocity T, respectively. The time series of beat candidate points is denoted as $\{b_j\}$, $j=1,2,...,N$, and the formula for calculating the correlation significance s_j^b is:

$$s_j^b=r_{fl,T}^{k}(b_j)+r_{ch,T}^{k}(b_j) \tag{22}$$

Among them,

$$r_{c,T}^{k}(k)=\frac{\sum_i r_{c,T}^{i}}{\max\left(r_{c,T}^{i}(s)\right)},c=\{ch,fl\} \tag{23}$$

Generally, the distance function is selected as the detection function for whether two points are beat points. There are various methods for calculating the distance between two points, most of which use $L-1$, $L-2$ norms and their deformations. Obviously, based on the relationship between beat points and non-beat points, the distance between two beat candidate points must increase as their time interval

deviates from the target cycle, and must decrease as the significance level of the next candidate beat increases. In this article, the distance between beat points b_i and b_j is set as follows:

$$\mathrm{d}\left(b_i, b_j\right) = \gamma d_T\left(b_i, b_j\right) - (1-\gamma) s_j^b \tag{24}$$

Among them, the calculation method for $d_T\left(b_i, b_j\right)$ is as follows:

$$d_T\left(b_i, b_j\right) = 1 - \exp\left\{-\frac{1}{\sigma^2}\ln^2\left(\left(b_i - b_j\right)/\tau_T\right)\right\} \tag{25}$$

Parameter $\gamma \epsilon (0,1)$ is used to control the weights of two variables in the distance function, while σ represents the degree to which the difference between the inter beat distance and the target period τ_T affects $d\left(b_i, b_j\right)$. In order to accurately detect the beat when the rhythm of music changes, the rhythm value needs to be taken into account in the distance between beat points. After obtaining the rhythm value T of a piece of music, find the significant peak of each frame s around the obtained rhythm value T. These peaks form a new rough rhythm value curve t_c, that is:

$$t_c(s) = \arg\max_{(1-\beta)T<t<(1+\beta)T}\left\{TG^{fl}(t,s) + TG^{ch}(t,s)\right\} \tag{26}$$

Among them, β represents the Changshu factor. Subsequently, the value of Q in interval $\left[\min\left(t_c\right), \max\left(t_c\right)\right]$ is taken as 4, and the rhythm value is estimated again. At this time, the formula for calculating the significance of beat points also needs to be updated, which is:

$$\dot{r}_c(k) = \frac{\sum r^i_{c,T(b_j)}}{\max_s\left(\sum_i r^i_{c,T(s)}(s)\right)} \tag{27}$$

Among them, $T(s)$ represents the maximum rhythm value corresponding to frame s. Therefore, rewrite equation (25) as follows:

$$d_T\left(b_i, b_j\right) = 1 - exp\left\{-\frac{1}{\sigma^2} ln^2\left(\left(b_i - b_j\right)/\tau_{T(b_j)}\right)\right\} \tag{28}$$

Based on the formula of inter beat distance, analyse the beat points as follows:

If $\left\{b_j\right\}, l \in L \subseteq \{1,...,N\}$ is the target beat sequence, then the optimal beat sequence $\left\{b_l^*\right\}$ should minimise the objective function, which can be expressed using the beat to beat distance formula as:

$$0\left(b_l^*, 1 \in L\right) = \sum_{1\in L}\mathrm{d}\left\{b_{i-1}^*, b_i^*\right\} \tag{29}$$

Using $C^*\left(b_i\right)$ to represent the minimum cost to reach beat point b_i, establish a recursive formula for dynamic programming:

$$C^*\left(b_i\right) = \min_s\left\{d\left(b_k, b_i\right) + C^*\left(b_k\right)\right\}, i = 1,2,...,N \tag{30}$$

$$\mathrm{path}\left(b_i\right) = \arg\min_{b_k}\left\{d\left(b_k, b_i\right) + C^*\left(b_k\right)\right\} \tag{31}$$

Among them, $\mathrm{path}\left(b_i\right)$ represents the optimal path taken by the previous beat point to reach b_i.

To obtain the optimal sequence, select a subset $\mathrm{C}\left\{\dot{b}_i\right\}$ of possible beat point positions at the end of the music clip, and first determine the last beat point as:

$$b_k = \arg\min_{b_k}\left\{C^*\left(\dot{b}_m\right)\right\} \tag{32}$$

By backtracking to find the optimal beat sequence, it can be expressed as:

$$\dot{b}_{i-1} = \mathrm{path}\left(b_l\right), l = K,...,2 \tag{33}$$

Based on the above design, the relative position of the beat points can be determined to complete the extraction of pop music singing beats.

This study innovatively constructed a music beat extraction framework that integrates time-frequency domain features, and achieved speed adaptive processing through feature map matrix and time bending technology. Creatively proposed a binary saliency model based on harmonic and impulse components, established a global periodic vector analysis mechanism and designed a dynamically updated rhythm value estimation method. Introducing inter beat distance function and adaptive weight adjustment strategy in the beat tracking stage, combining the target cycle deviation with significance level and implementing optimal beat sequence search through dynamic programming algorithm. This method breaks through the limitations of traditional beat detection in terms of adaptability to speed changes, achieves collaborative optimisation of feature analysis and beat tracking and provides new ideas for beat extraction in complex music environments

4 Experiments and analysis

4.1 Experimental environment settings

In order to verify the progressiveness of the proposed method, experimental research was carried out. The environmental parameter settings for this experiment are shown in Table 1.

Table 1 Experimental environment parameter settings

Project	*Index*
Operating system	Windows 11
Processor	Intel Core i7 - 12700H @ 2.30GHz
Memory	32GB DDR5 4800MHz
Programming language	Python 3.9
Audio processing library	Librosa 0.10.0, PyDub 0.25.1

In the above experimental environment, 50 popular Chinese pop songs were collected recently, and a Chinese pop song beat annotation data set was carefully created. The annotation information of this data set covers the beat and beat information of the songs, aiming to provide comprehensive and accurate evaluation basis for beat extraction algorithms. The annotation data is saved in the form of two column sequences. The first column is the timestamp of the beat, accurate to milliseconds, used to identify the specific position of the beat in the audio. The second column is the section number and beat number where the beat is located. If the beat number is '*x*.l' (where *x* is the section number and l is the beat number identifier), it indicates that the position is a section line, that is, a beat. The annotation accuracy is set to 50ms. To ensure the accuracy and reliability of annotated data. The total audio duration of the data set selected this time is 3 hours and 47 minutes, and the labelling error is strictly controlled within 50 ms, fully ensuring the quality of the data set. In terms of song selection, the 50 songs in the data set are all Chinese pop styles, with a release time span from 2000 to 2025, covering the characteristics of popular music in different periods. In terms of singers, it includes 22 songs sung by female singers, 26 songs sung by male singers and 2 songs sung by male and female singers, with a certain degree of diversity and representativeness. Based on this, conduct experimental testing.

The training sample data set constructed in this study includes 320 kbps MP3 audio files of 50 Chinese pop songs, with an average duration of 4 minutes and 33 seconds per song. The audio sampling rate is uniformly 44.1 kHz, with a quantisation accuracy of 16 bits and adopts a stereo dual channel format. The data set contains a total of 12,850 annotated beat points, of which retake points account for 23.6%. The training set and test set are divided in a 7:3 ratio. The training set contains 35 songs with a total of 8995 beat samples, while the test set consists of 15 songs with 3855 beat samples. All audio files have been manually annotated by professional audio engineers, with a beat timestamp annotation error controlled within ± 25 ms.

In experimental indicator design, functional validation mainly focuses on filtering effectiveness indicators, including feature retention, noise suppression rate and auditory lossless standards. The core indicators for rhythm extraction are divided into basic accuracy indicators and robustness supplementary indicators. The basic accuracy indicators include rhythm detection accuracy, missed detection rate and false detection rate, while the robustness supplementary indicators include time-varying tolerance and complexity adaptation indicators. In order to fully demonstrate the performance of the proposed method, a comparative test was conducted between the proposed method and the Long et al. (2024) and Yang (2024) method.

4.2 Experimental results

Firstly, conduct functional validation of the proposed method and select the beat extraction method to pre-process popular music signals with five different rhythms. The filtering results are shown in Figure 2.

Figure 2 Filter processing result

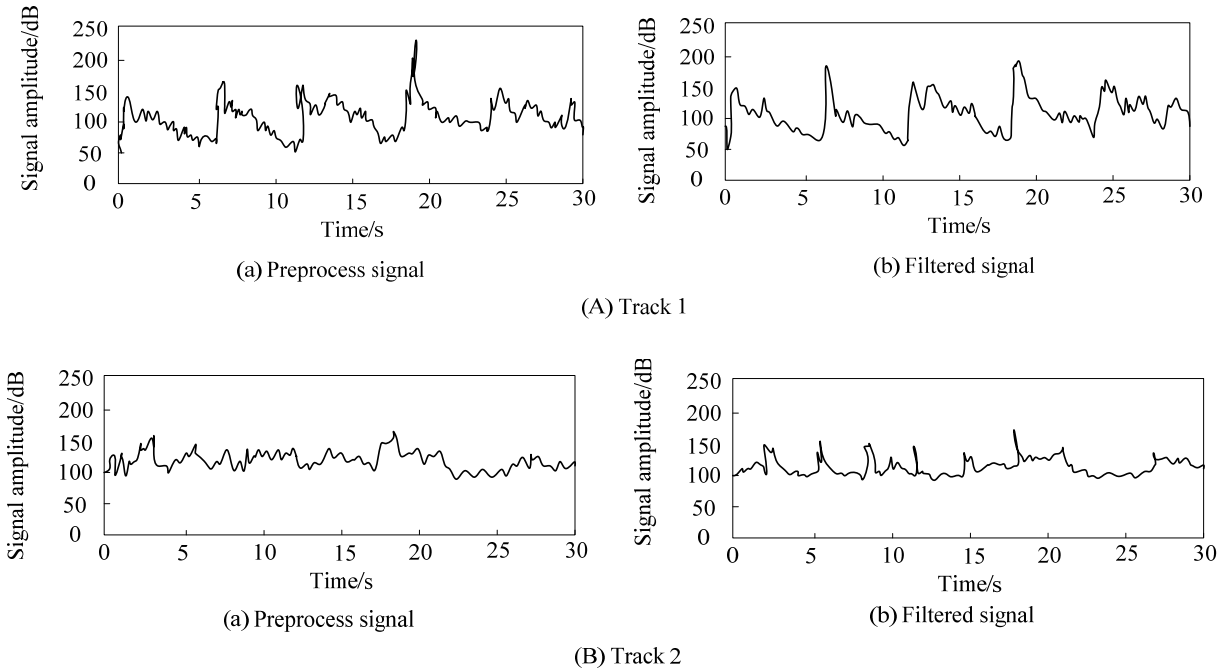
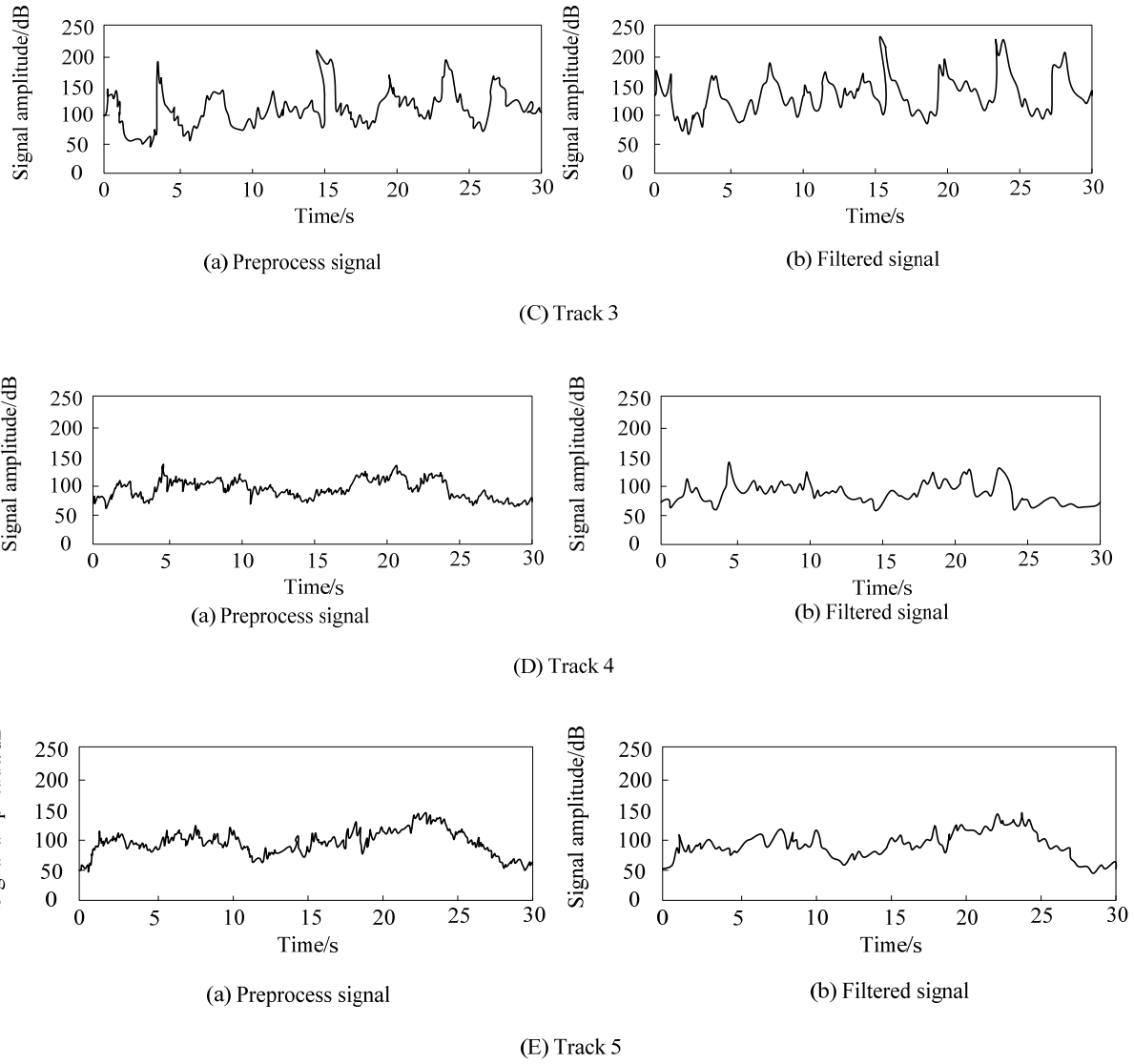


Figure 2 Filter processing result (continued)



As shown in Figure 2, the proposed method can effectively remove noise and preserve the main characteristics of music signals, whether for music with strong rhythm or music with complex rhythm changes and fast regular changes. The application effect is good.

Next, continue to apply the proposed method for rhythm extraction, and the results are shown in Table 2 below.

Table 2 Effect of rhythm extraction

Track number	*Length of music/s*	*Actual number of beats /piece*	*Detect the number of beats/piece*	*Miscount the number of beats/piece*	*Number of missed beats/piece*
1	30	14	14	0	0
2	30	15	14	0	1
3	30	13	13	0	0
4	30	13	13	0	0
5	30	14	14	0	0

As shown in Table 2, the proposed method demonstrated high accuracy in beat extraction, with the detected beat count being consistent with the actual beat count. The number of false and missed beats were both at a low level, indicating its ability to effectively identify beat positions in music signals and accurately extract beat information; From the analysis of each track, it can be seen that tracks 1, 3, 4 and 5 have exactly the same number of detected beats as the actual ones, with both false and missed beats being 0. This indicates that for tracks with relatively stable rhythms, the proposed method can perfectly extract all beat information, while track 2 missed one beat. Considering the length of the music and the actual number of beats, this error is within an acceptable range. It may be due to some subtle changes in the music rhythm that the proposed method failed to accurately identify the beats, and the overall recognition effect is good.

Finally, the Long et al. (2024) and Yang (2024) method were selected as the comparison methods, and the missed detection rate and false detection rate were used as performance evaluation indicators to conduct comparative experiments. The results are shown in Figures 3 and 4.

Figure 3 Comparison of missed detection rates among three methods (see online version for colours)

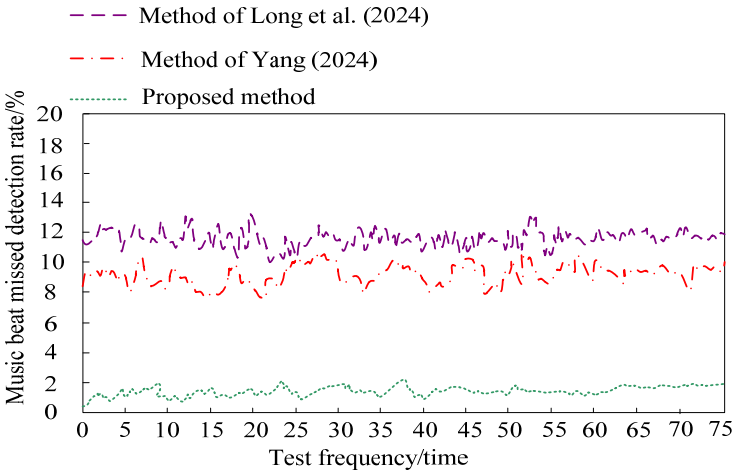
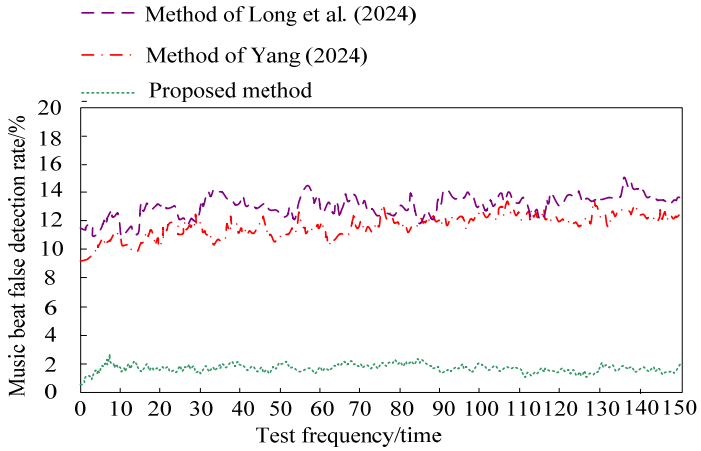


Figure 4 Comparison of false detection rates among three methods (see online version for colours)



From the analysis of Figures 3 and 4, it can be seen that the proposed method has a missed detection rate of only 2.1% and a false detection rate of only 2.5%. However, using both comparison methods, the missed detection rate is higher than 7% and the false detection rate is higher than 9%. This proves that the proposed method can accurately capture beats in music signals, with only a very small number of beats being missed. At the same time, the occurrence of incorrectly identifying non beat positions as beats is also very rare. It has higher accuracy and reliability in beat extraction tasks, and the application effect is good. This is because the method proposed in this article improves signal quality through discretisation and normalisation pre-processing, combines time-frequency domain joint analysis to comprehensively capture music features, and adopts a multi feature fusion strategy to effectively integrate beat cycle and inter beat distance information, making beat detection more accurate. This comprehensive feature extraction and analysis method significantly reduces noise interference, enhances the robustness of beat localisation and achieves significantly better performance than the comparison methods in terms of missed detection rate and false detection rate.

5 Conclusion

Against the backdrop of rapid development in music information processing technology, this study innovatively proposes an audio analysis method based on multifeature fusion to address the key challenges faced by pop music singing rhythm extraction, such as environmental noise interference and complex instrument accompaniment. This method constructs a feature map matrix through a time-frequency domain joint analysis framework, cleverly introduces time bending technology to achieve speed adaptation, and innovatively establishes a binary saliency model to optimise rhythm value estimation. The experimental data-driven research confirmed the excellent performance of the method, achieving a low missed detection rate of 2.1% and a low false detection rate of 2.5% in beat detection tasks, significantly better than existing methods. Based on these innovative discoveries and experimental verification, this study has opened up new technological paths for popular music information processing. In the future, we will continue to optimise model robustness, expand the application boundaries of methods in music creation and analysis, and inject new impetus into the development of music technology.

Acknowledgements

This work was supported by 2019 Heilongjiang Provincial Art and Science Planning General Project, Thinking and Construction of Integrated Teaching of Pop Music and Dance Course, Project No.: 2019B130; 2017 Heilongjiang Provincial Art and Science Planning General Project, Research on Contemporary Network Pop Music and Its Development Trend, Project No.: 2017B131; 2024 Heilongjiang Provincial Philosophy and Social Science Research Planning Support and Joint Construction Project, Value implications and Practice Path of Northeast Anti-Japanese United Front Opera in Heilongjiang Province, Project No.: 24YSE005; National Social Science Fund Cultivation Project of Jiamusi University, Research on the protection and inheritance of songs and ballads of the Northeast Anti-Japanese Union, Project No.: JMSUGPRW2308.

Declarations

All authors declare that they have no conflicts of interest.

References

Dong, L.S. (2022) 'Optimization simulation of dance technical movements and music matching based on multifeature fusion', *Computational Intelligence and Neuroscience*, No. 27, pp.8679748–8679757.

Fu, D.H. (2024) 'Non-stationary noise filtering of 4K high-definition all-media broadcasting vehicle audio signal', *Microcomputer Applications*, Vol. 40, No. 6, pp.50–52+64.

Guo, K.L. and Wang, J.Y. (2024) 'Audio signal endpoint detection system in non-stationary strong noise environment', *Modern Electronics Technique*, Vol. 47, No. 10, pp.18–22.

Han, B.B., Cheng, K. and Wang, Y.J. (2023) 'Multi-feature fusion music classification algorithm based on CGABC-SVM', *Computer and Digital Engineering*, Vol. 51, No. 4, pp.820–825.

Han, D.H., Kong, Y.R. and Meng, Z.Y. (2024) 'Research on emotion recognition method of music multimodal data', *Journal of Northeastern University (Natural Science)*, Vol. 45, No. 6, pp.776–785+792.

Jing, P.X. (2024) 'Research on music fountain control system based on feature extraction and BP neural network', *Automation and Instrumentation*, Vol. 14, No. 7, pp.313–316.

Lan, C.F., Jiang, P.W. and Chen, H. (2024) 'Multi-head attention time domain audiovisual speech separation based on dual-path recurrent network and Conv-TasNet', *Journal of Electronics and Information Technology*, Vol. 46, No. 3, pp.1005–1012.

Li, Y.A. (2024) 'Piano audio signal recognition algorithm based on multi-dimensional spectrum diagram', *Techniques of Automation and Applications*, Vol. 43, No. 3, pp.169–171+176.

Liu, F. (2024) 'Automatic control of dance robot based on voice print emotion feature extraction', *Automation and Instrumentation*, Vol. 11, No. 9, pp.276–279+284.

Long, E.H., Wang, G. and Mo, L.F. (2024) 'An audio recognition method based on SOM and spiking neural network', *Chinese Journal of Sensors and Actuators*, Vol. 37, No. 11, pp.1885–1892.

Potempski, F., Sabo, A. and Patterson, K.K. (2022) 'Technical note: quantifying music-dance synchrony during salsa dancing with a deep learning-based 2D pose estimator', *Journal of Biomechanics*, Vol. 3, No. 14, pp.1–7.

Wang, Y. (2023) 'Intelligent auxiliary system for music performance under edge computing and long short-term recurrent neural networks', *Plos One*, Vol. 18, No. 5, pp.1–20.

Wu, K.Y., Ruan, W.D. and Zhou, D.F. (2023) 'Syllable clustering analysis-based passive acoustic monitoring technology and its application in bird monitoring', *Biodiversity Science*, Vol. 31, No. 1, pp.126–136.

Yang, J. (2023) 'Research on anti-interference of full-band music signal based on cyclic spectrum characteristics', *Information Technology*, Vol. 21, No. 5, pp.121–125+130.

Yang, L.Y. (2024) 'Design of music beat recognition system based on artificial intelligence technology', *Techniques of Automation and Applications*, Vol. 43, No. 3, pp.128–131.

Yang, S.Y. and Li, X. (2023) 'Lightweight end-to-end architecture for streaming speech recognition', *Pattern Recognition and Artificial Intelligence*, Vol. 36, No. 3, pp.268–279.

Yu, M., Liu, Z.W. and Shi, S. (2023) 'Audio-video emotion recognition based on residual network and coarse-fine granularity', *Computer Engineering and Design*, Vol. 44, No. 7, pp.2192–2199.

Zhao, X.F., Peng, X. and Chang, Y.N. (2023) 'Robotic dance automatic generation system based on voiceprint emotion analysis', *Modern Electronics Technique*, Vol. 46, No. 15, pp.84–88.

Zhong, Z.P., Wang, H.L. and Su, G.B. (2023) 'Music emotion recognition fusion on CNN-BiLSTM and self-attention model', *Computer Engineering and Applications*, Vol. 59, No. 3, pp.94–103.

Zhu, Y. (2022) 'Recognition method of matching error between dance action and music beat based on data mining (Retracted Article)', *Security and Communication Networks*, pp.1–8.