ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Dynamic Visualization Method of Vocal Melody Based on Audio Feature Extraction and Image Processing

Dihua Zhang[1] *and Jia Kuang*[1,*]

[1] College of Music, Hengyang Normal University, Hengyang 421001, Hunan, China

SUMMARY: *With the rapid development of computer technology, the dynamic visualization of vocal music melody has attracted more and more attention. The dynamic visualization of vocal music melody is a way of expression and presentation of music. The dynamic visualization of vocal music melody is a way to express more music information in various forms with vision as the center and with music, which is a new form of cultural exchange in the information age. This paper focused on the dynamic visualization of vocal melody, and proposed audio feature extraction based on auto correlation function (ACF) and improved ACF algorithm. The improved ACF algorithm was more efficient in extracting audio features, which was also conducive to the dynamic visualization of vocal music melody. The experimental results in this paper showed that the extraction time of ACF algorithm and improved ACF algorithm on MIR-1K dataset was 10.8s and 3.6s respectively when the pitch was 180, and the extraction time of ACF algorithm and improved ACF algorithm on MedleyDB dataset was 12.3s and 2.8s respectively when the pitch was 180. It can be found that the extraction time of the improved ACF algorithm was less than that of the ACF algorithm, which also showed that the improved ACF algorithm had higher extraction efficiency.*

KEYWORDS: *Visualization of Vocal Melody, Auto-correlation Function, Audio Feature Extraction, Image Processing*

1 Introduction

Music visualization has moved from simple spectrum animation to feature-aware, interaction-oriented, and crossmodal design. Recent studies have shown that melody, timbre, dynamics, and user interaction can be mapped to visual variables such as color, geometry, motion, and spatial depth. For vocal music, however, the visual quality of the output still depends on whether the system can track the main melody accurately enough to support frame-level visual updates.

The core technical difficult point is not the rendering itself, but the stability of melody picking out under accompaniment disturbance, pitch undulation, and cross-dataset change. In the case that pitch salience has an abrupt change, or the estimation of local features is not steady, therefore the animation that is obtained will become not continuous, visually repetitive, or semantically not strong. This therefore lets vocal-melody picture-showing become a together problem of sound character taking-out, time building-model, and picture changing, not a purely picture problem.

This study therefore focuses on an improved ACF-based feature extraction scheme for

*hysfxykj001@126.com
https://doi.org/10.65102/is2026345 Published: 30 April, 2026

dynamic visualization of vocal melody. The contribution is threefold. First, the manuscript reorganizes the method as a verifiable pipeline linking short-time analysis, pitch estimation, similarity calculation, and visual mapping. Second, it evaluates the method on MIR-1K and MedleyDB using extraction rate, error rate, classification accuracy, and processing time. Third, it clarifies how the extracted features can support educational and performance-oriented applications in singing analysis and feedback. In contrast to descriptive visualization papers that mainly emphasize rendering style, the present study keeps the signal-processing stage explicit so that each visual response can be traced back to a measurable audio descriptor. This point is important for academic writing because the quality of the visualization is treated here as an outcome of feature reliability rather than as an isolated design effect.

2 Related Work

The currently existing research works about music visualization can be separated into three parts: feature to visual mapping, interaction design, and immersive presentation. The investigation which is made by Lima et al. [1] has the result that many systems still depend on a limited number of low-level audio characteristics [2], while more new research has investigated timbre-form correspondence, architecture [3] of real-time visualization [4] pushed by audio, and dynamics visualization that faces practice. These research works verify that visualization quality is decided by both the relatedness of the picked features and the understandability of the visual mapping.

A second line of work extends music visualization into mixed-reality and AI-assisted environments. Erdmann et al. [5] used real-time audio analysis and crossmodal correspondences to build a mixed-reality concert visualization, while Huang et al. [6] designed an AI-driven system that generates meaningful audio-responsive visuals in real time. Yu [7] further reported that deep neural networks can improve the discriminability and responsiveness of visualization models. Putting together, these research works point out that visualization systems can get benefits from richer feature collections and more stable time-domain control. They also point out that modern visualization study already is not restricted within the decorative frequency spectrum animation; on the contrary, it more and more relies on the mappings which are driven by data, the logic of interaction, and the organization of features which have meaning in perception.

For the extraction of vocal melody, the recent advancement has moved from heuristic pitch tracking toward deep or graph-based model construction. RMVPE [8] promotes the robust estimation of vocal pitch in polyphonic music, graph modeling methods [9] increase the explainability in melody salience estimation, and networks which are harmonic-aware or transformer-based [10, 11] improve the modeling of long-range dependency. Jing and other persons [12] further carry out the integration of attention aggregation and self-consistency training in the work of singing melody extraction. These methods make clear that the accurate estimation of melody still is the key precondition for the rendering that has meaning in visual aspects.

The current gap is that many visualization studies emphasize rendering strategies, whereas many melody-extraction studies emphasize recognition accuracy, and the connection between the two is often weak. In particular, the literature seldom explains how a specific pitch-extraction strategy improves the continuity, responsiveness, and interpretability of dynamic visual output in vocal music. This paper addresses that gap by using a comparative ACF/improved-ACF framework and by relating extraction performance directly to visualization-oriented requirements such as stability, efficiency, and multi-feature control.

3 Methods Based on Audio Feature Extraction and Image Processing

3.1 Dynamic Visual Design of Vocal Music Melody

Music has always been an effective tool to express emotions. However, as traditional music listening no longer meets people's desire for a colorful world, dynamic visualization of vocal music melody has become a feasible choice. The dynamic visualization of vocal music melody also coincides with the rapid development of music in image processing, virtual reality, digital signal processing and other fields. Therefore, the dynamic visualization technology of vocal music melody has attracted researchers' interest and found practical use in other fields. Dynamic visualization of vocal music melody enriches the auditory experience by naturally integrating auditory and visual systems, and provides services for a wide range of users. With the improvement of living standards, people are more and more interested in spiritual and cultural life. People hope to observe more diversified visualization effects, and realize how different dynamic visualization effects of vocal music melody vary according to the type of music they listen to, so as to better understand the music theme when listening to dynamic visualization of vocal music melody.

At present, most of the dynamic visualization of vocal music melody is based on a single feature, and this visual effect cannot meet the needs of today's people. Therefore, this paper proposes audio feature extraction and image processing to enrich the dynamic visualization of vocal music melody, and facilitates the realization of 3D image rendering. The dynamic visualization design of vocal music melody is shown in Figure 1.

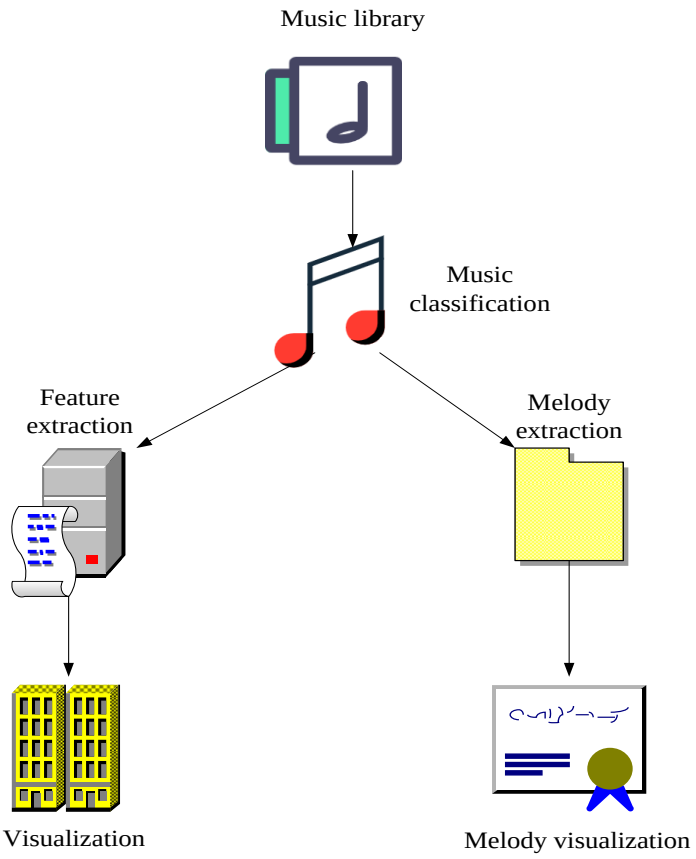


Figure 1: Dynamic visual design of vocal melody

As shown in Figure 1, the dynamic visualization of vocal music melody first classifies music from the music library, then extracts features, and finally performs visualization processing. Because most of the visualization effects are based on a single feature, a dynamic visualization method of vocal music melody based on multiple features is adopted. The music in the music library is screened with multiple characteristics. Then, according to the specific requirements of visualization processing, a variety of music forms of visual design has been achieved.

The dynamic visualization process of vocal music melody mainly takes the content of sound as a non subjective expression, which includes the interpretation and judgment of sound, laying the foundation for future analysis [13].

The use of visualization technology can transform the vocal music melody from an audible information to a visible information. The dynamic visualization of vocal music melody has important application value in digital entertainment, disabled teaching and other aspects. At the same time, with the progress of science and technology, the dynamic visualization of vocal music melody can provide more convenience for human beings [14, 15]. Dynamic visualization of vocal melody is shown in Figure 2.

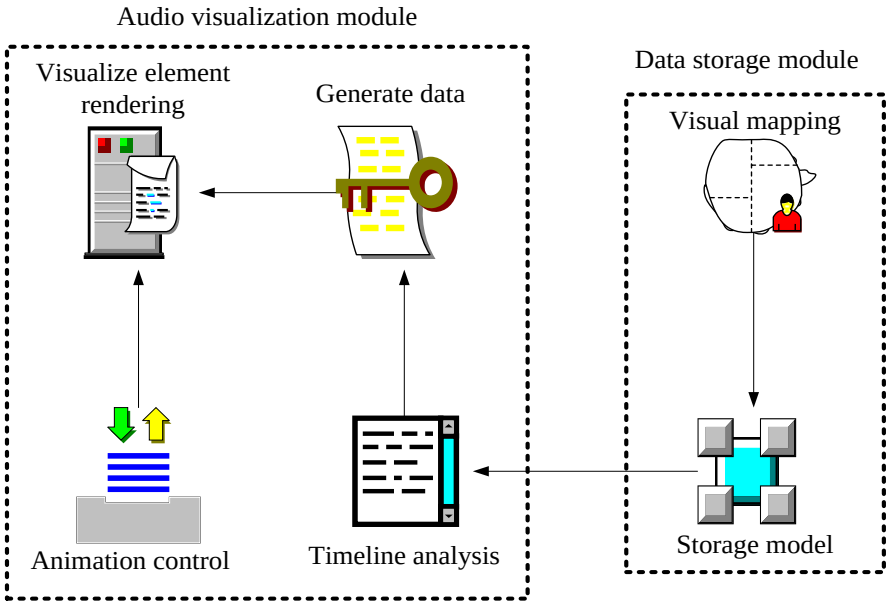


Figure 2: Dynamic visualization of vocal melody

As shown in Figure 2, the dynamic visualization of vocal music melody includes a data storage module and an audio visualization module. The audio visualization module is an intuitive visualization component, which first performs time analysis and generates data according to the sound file itself. Then, it overlays the animation control to get the final dynamic effect. In this process, the data of the animation control module is generated through the user's interaction, and the visual elements also change with the user's actions. Users can simply upload the content to be displayed on the Internet and hand over a large number of computing processes to the computer.

3.2 Dynamic Visualization of Vocal Music Melody Based on Image Processing

The visual effect of dynamic visualization of vocal music melody is expressed in the form of images, which are inseparable from the translation, zoom, rotation, perspective, wave

transformation and spherical transformation of images. This processing method can produce a variety of visual effects with strong visual impact. All objects in the scene can be visualized. Some objects can be displayed in the whole process of visualization, while others can be dynamically displayed in a specific time. It is very difficult to create a complex and beautiful scene with simple objects, which requires finding sufficiently complex image objects and conducting appropriate image processing. Limited parameters are used to control visual effects. The application of dynamic visualization of vocal music melody is shown in Figure 3.

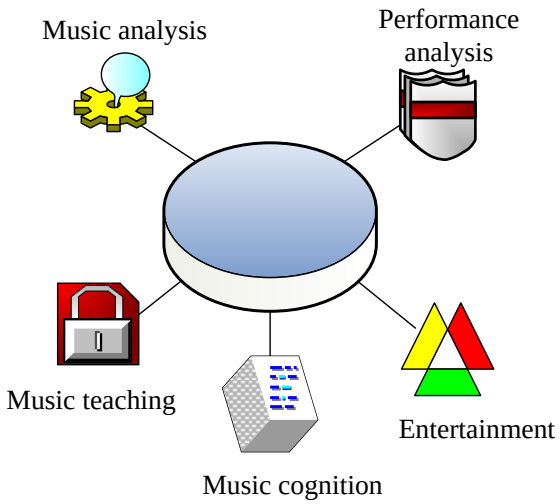


Figure 3: Application of dynamic visualization of vocal melody

As shown in Figure 3, dynamic visualization of vocal music melody has a wide range of applications, such as music analysis, performance analysis, music teaching, music cognition and game entertainment. Vocal music conveys human thoughts and feelings in an auditory way, which is an important ideology of social reality. However, the visual stimulus is stronger than the auditory stimulus and can leave a deep impression. Sometimes, the artistic conception of vocal music melody is difficult to comprehend. However, the combination of sound effects and pictures through image processing certainly makes people feel more vivid and can better understand the singer's mood. With the rapid development of image processing technology, the dynamic visualization of vocal music melody has also been developed.

The use of computers can basically achieve any image processing, which can meet the needs of most fields. The development of image processing technology has penetrated into many important fields. The application range of image processing technology is shown in Figure 4.

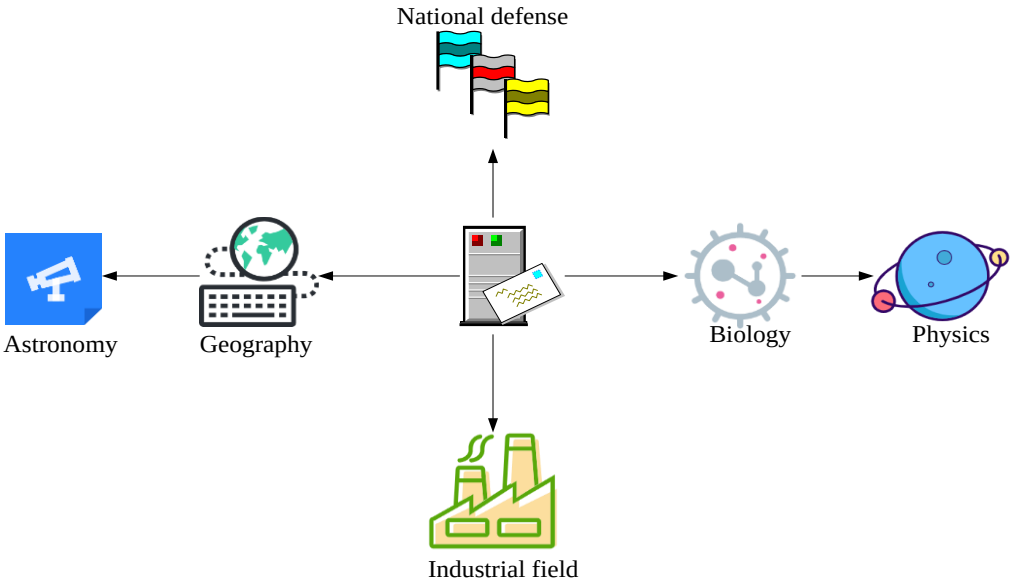


Figure 4: Application scope of image processing technology

As shown in Figure 4, image processing technology has achieved a lot in the fields of geography, physics, astronomy, biology, national defense and industry. However, up to now, there has not been a scientific and accurate definition of image features. This is because the problem of image processing involves too wide a scope, and the definitions of all aspects are different. It is difficult to give an accurate and unified definition. Generally speaking, image processing starts from the extraction of features in the image, and finally evolves into algorithm analysis related to image operation.

The advantage of image feature matching is that it can be applied to the matching of single or multiple patterns at the same time. However, due to the large number of feature extraction problems in the process, the work in target detection and segmentation is very difficult. In addition, in the process of manual or automatic feature extraction, there are some unrecoverable errors, which have a great impact on matching. Since this method does not need to extract feature points, it can directly use gray image registration.

At present, the simplest image similarity measure SSD is the square sum of density difference between images:

$$\varphi_{SSD} = -\frac{1}{n}\sum\left(\varphi_A(a) - \varphi_A(T(a))\right)^2 \tag{1}$$

Under the assumption of the same imaging mode, this measurement is carried out. When the noise is Gaussian, SSD is the best matching metric. Since the premise of similarity measurement is to assume that image models are consistent, this method can only be used for single pattern matching.

Under this assumption, the same imaging method can achieve alignment, but its limitations are also large. A common method is normalized cross-correlation. In this case, the similarity between images can be measured by normalized cross-correlation φ_{cc}:

$$\varphi_{cc} = \frac{\sum\left(\varphi_A(a) - \mu_A\right)\left(\varphi_A(T(a)) - \mu_B\right)}{\sqrt{\sum\left(\varphi_A(a) - \mu_A\right)^2\left(\varphi_B(T(a)) - \mu_B\right)^2}} \tag{2}$$

Among them, μ_A and μ_B correspond to the average gray value of the two images respectively. Although φ_{cc} is more flexible than SSD, there are still many limitations. An important part of this method is to combine the feature space of probability distribution (image gray) to accumulate the gray histogram of two images, so as to visualize the feature space. Given φ_A and φ_B images, the information between these two images is defined as:

$$I(\varphi_A,\varphi_B)=H(\varphi_A)+H(\varphi_B)-H(\varphi_A,\varphi_B) \tag{3}$$

Among them, $H(\varphi_A)$ and $H(\varphi_B)$ are the entropy of the image, and $H(\varphi_A,\varphi_B)$ is the joint entropy of the image. The position of two images in space is consistent. Therefore, the mutual information between them should be maximum, which is the principle that mutual information can be used as a similarity measure.

When calculating the similarity of local or short-term features of vocal music, the short-term features extracted each time can form a high-dimensional vector for similarity calculation. The common method to calculate the similarity between the eigenvectors A and B of sound is to calculate the Euclidean distance between the eigenvectors. The Euclidean distance of point (a_1,b_1) and point (a_2,b_2) can be expressed by the following formula:

$$d=\sqrt{(a_1-a_2)^2+(b_1-b_2)^2} \tag{4}$$

The Euclidean distance is extended to the minimum distance on two planes, and its properties are basically consistent with those of the Euclidean distance on two planes in three-dimensional space. The three-dimensional Euclidean distance between point (a_1,b_1,z_1) and point (a_2,b_2,z_2) in three-dimensional space can be expressed as follows:

$$d=\sqrt{(a_1-a_2)^2+(b_1-b_2)^2+(z_1-z_2)^2} \tag{5}$$

Therefore, the Euclidean distance can be extended to n-dimensional space, which is the shortest distance between two points, and can also be expressed by Euclidean distance:

$$d(A,B)=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2} \tag{6}$$

It should be noted that point A in n-dimensional space is a point in n-dimensional space, and A can be expressed as $(a_1,a_2,...,a_n)$. Among them, a_n is the n-th coordinate of point a. In this paper, the concept of Euclidean distance discriminant method is derived from n-dimensional Euclidean distance. First, the n-dimensional feature vector is extracted, and it is used for Euclidean distance operation with the n-dimensional feature vector extracted during training. Euclidean distance has the characteristics of simple concept, easy realization and high recognition rate.

3.3 Dynamic Visualization of Music Based on Audio Feature Extraction

From the perspective of vocal melody, low-grade sound quality includes amplitude, frequency, phase, etc. The higher level of pitch and strength are the characteristics of the scale. Good visual effect of vocal music melody needs to extract higher level music features, while it is difficult to extract higher level sound quality features. The current visual effects are mostly

achieved by changing the low-grade sound quality.

Each feature extraction is time-consuming, unrealistic and meaningless. Therefore, in order to shorten the running time of extraction and meet the real-time requirements, it is necessary to extract those important music features according to the requirements of the dynamic visualization application of vocal music melody, while ignoring or removing those unnecessary music features. The categories of important musical features are shown in Figure 5.

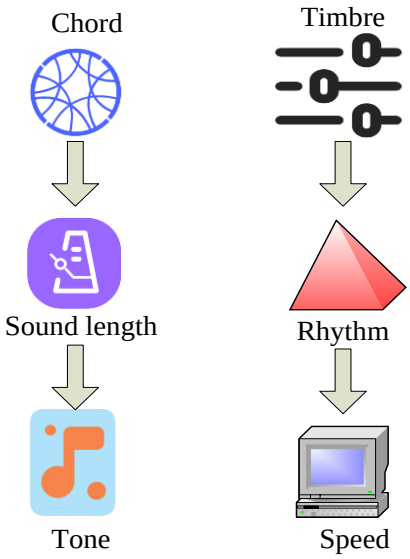


Figure 5: Types of important music features

As shown in Figure 5, there are various features in music, such as timbre, pitch, length, chord, speed, beat and some new features proposed recently. Audio signal can be visualized in time and frequency domain, showing amplitude state in time range and energy in frequency area.

Audio feature extraction must be used in music feature analysis, and the extracted features are different according to the requirements. The feature extraction of hard disk is mainly in time and frequency. Because their characteristics cannot be read directly from the sound, they need to be preprocessed.

(1) ACF based audio feature extraction

Audio feature extraction is to convert the voice signal to a certain function, and then extract the tone sequence from the converted voice signal. The transformation function mainly includes ACF. The cycle value is obtained by analyzing the extreme points of the converted waveform, and the corresponding frequency is calculated.

The calculation process of ACF is shown in Formula (7):

$$r_t = \sum_{i=0}^{L-t-1} a_i \cdot a_{i+t} \tag{7}$$

After ACF transforms cycle signal a_i, the function waveform reaches its highest value at the integral multiple of its cycle. The offset corresponding to the extreme value in the function waveform with the maximum value is used as the audio cycle value a_{i+t}.

In audio feature extraction, people often think of using time as information feature extraction, that is, analyzing the time-domain waveform of audio signal, which is called

time-domain feature extraction.

The short-term average energy of audio signal is also a characteristic quantity, which can reflect the amplitude:

$$E_n = \sum_{m=-\infty}^{\infty} [a(m) \cdot w(n)]^2 * h(n) \tag{8}$$

In the formula, E_n represents the short-time energy starting from point n of the serial number, and $w(n)$ is the window function.

The short-time average zero crossing rate can roughly reflect the average frequency of the signal in the short-time audio signal frame. Therefore, the frequency domain characteristics can be calculated by the following formula:

$$Z_n = \frac{1}{2} \sum_{-\infty}^{\infty} |\mathrm{sgn}[a_N(m) - \mathrm{sgn}[a_N(h)]]| \tag{9}$$

The m-th signal value in the audio signal frame is represented by $a_N(m)$, and the function window length is represented by $a_N(h)$. Because the frequency can be estimated by using the short-time zero crossing rate, there is a relationship between them. If the current frame of the audio has a higher zero crossing rate, the frequency value is higher, and vice versa.

In order to extract audio features in the frequency domain, waveform signals in the time domain must be converted into signals in the frequency domain. At present, Fourier analysis is the most widely used and the best effect. If the input signal is e^{-jwm}, the signal is subjected to short-time Fourier conversion $A_n(e^{jw})$:

$$A_n(e^{jw}) = \sum_{m=-\infty}^{\infty} a(m) w(n-m) e^{-jwm} \tag{10}$$

After the short-time Fourier transform, because the short-time power spectrum is the result of the Fourier transform amplitude, the most direct audio feature can be obtained, that is, the short-time power spectrum $P_n(e^{jw})$:

$$P_n(e^{jw}) = \left| A_n(e^{jw})^2 \right| = \sum_{m=-\infty}^{\infty} R_n(k) e^{jwm} \tag{11}$$

The short-time power spectrum is the Fourier transform of short-time autocorrelation function, so that the characteristics of sound can be extracted from the frequency domain $R_n(k)$.

(2) Audio feature extraction based on improved ACF

In the dynamic visualization of vocal music melody, different shapes, sizes, colors, animations and other forms are used to achieve the unity of vision, hearing and psychology, and make the vocal music melody visualized and intuitive. At the same time, when visualizing multiple vocal melodies, it is also necessary to analyze the correlation between vocal melodies. Dynamic time warping path length and Euclidean distance of musical tonality are used to visually display the correlation.

The improved ACF introduced in this paper is a new method to extract vocal music

sequences from audio digital signals, which can process more complex vocal music melody data. The improved ACF algorithm uses continuous hearing based on time domain and pitch to establish pitch continuity on the basis of calculating pitch significance. In real life, the basic frequency of vocal music melody changes with time. The basic frequency detection of pitch is to convert the vocal melody signal into a function. Then the analytic judgment method is used to detect the repetitive waveform from it, so as to obtain the pitch period.

The specific process of improving ACF algorithm is shown in Formula (12):

$$s = \frac{t}{LM} - 1 \tag{12}$$

Among them, t represents the frame length of the digital signal, and LM represents the frame shift. In order to determine the tone period, the original sound signal must be converted into a function waveform that is easier to detect the period. This is the basic stage of tone estimation using the improved ACF algorithm:

$$d_t = \frac{1}{L-t} \sum_{i=0}^{L-t-1} |a_i - a_{i+t}| \tag{13}$$

Among them, t represents the periodic waveform data to be calculated. The improved ACF algorithm transformation first defines the sgn indicator function. At the same time, it defines the total number SY_t of data points with two waveform related data values as:

$$SY_t = \sum_{i=0}^{L-t} \mathrm{sgn}(a_i) \tag{14}$$

MX_t and MN_t respectively represent the sum of the maximum and minimum values of the data corresponding to the two waveforms:

$$MX_t = \sum_{i=0}^{L-t} \mathrm{sgn}(a_i, a_{i+1}) \cdot Max(|a_i|, |a_{i+1}|) \tag{15}$$

$$MN_t = \sum_{i=0}^{L-t-1} [\mathrm{sgn}(a_i, a_{i+1}) \cdot Min(|a_i|, |a_{i+1}|)] \tag{16}$$

In the formula, $(|a_i|, |a_{i+1}|)$ represents the absolute value operation of real number. After obtaining the signal waveform after improved ACF conversion, the periodic value is detected. Tone is an important feature of vocal music, and its extraction accuracy directly affects the visualization of vocal melody. The accuracy of pitch estimation is not high, and the interference of sound quality and harmonic components is large. Therefore, a higher pitch value can be obtained by using improved ACF for pitch estimation.

4 Comparison Experiment of Audio Extraction Effects of Different Algorithms

The simulation experiment run under the software environment of Windows 10. This paper

used MIR-1K dataset and MedleyDB dataset as the experimental dataset.

4.1 Accuracy of Audio Extraction with Different Algorithms

First, 60 songs were selected as the basic database. The main tracks and consonants were recorded, and their features were archived. 60 pieces of music were used as training data for training.

In this paper, a series of comparative experiments were carried out to verify the effectiveness of the improved ACF algorithm. First, ACF algorithm and improved ACF algorithm were used to extract 60 music pieces respectively. The extraction rate of ACF algorithm is shown in Table 1.

Table 1: Extraction rate of ACF algorithm

Number of music	Successfully extracted	Extraction rate
10	4	40%
20	10	50%
30	18	60%
40	26	65%
50	36	72%
60	45	75%

As shown in Table 1, when ACF algorithm extracts 10 pieces of music, 4 pieces of music are extracted, and the extraction rate is 40%. When 20 pieces of music are extracted, 10 pieces of music are extracted, and the extraction rate is 50%. When 30 pieces of music are extracted, 18 pieces of music are extracted, and the extraction rate is 60%. When 40 pieces of music are extracted, 26 pieces of music are extracted, and the extraction rate is 65%. When 50 pieces of music are extracted, 36 pieces of music are extracted, and the extraction rate is 72%. When 60 pieces of music are extracted, 45 pieces of music are extracted, and the extraction rate is 75%

The extraction rate of improved ACF algorithm is shown in Table 2.

Table 2: Extraction rate of improved ACF algorithm

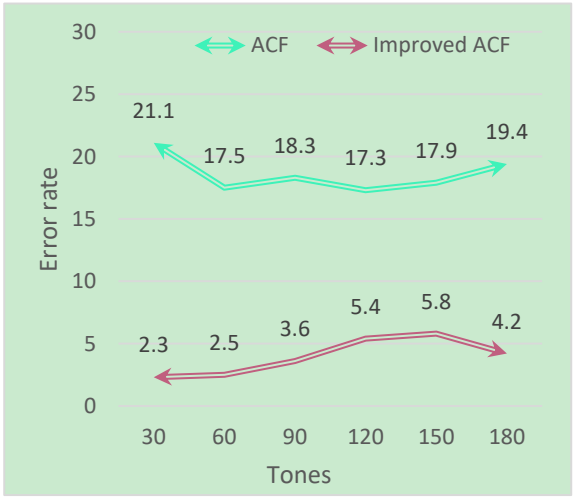
Number of music	Successfully extracted	Extraction rate
10	6	60%
20	15	75%
30	24	80%
40	34	85%
50	45	90%
60	57	95%

As shown in Table 2, when the improved ACF algorithm extracts 10 pieces of music, 6 pieces of music are extracted, and the extraction rate is 60%. When 20 pieces of music are extracted, 15 pieces of music are extracted, and the extraction rate is 75%. When 30 pieces of music are extracted, 24 pieces of music are extracted, and the extraction rate is 80%. When 40 pieces of music are extracted, 34 pieces of music are extracted, and the extraction rate is 85%. When 50 pieces of music are extracted, 45 pieces of music are extracted, and the extraction rate is 90%. When 60 pieces of music are extracted, 57 pieces of music are extracted, and the extraction rate is 95%.

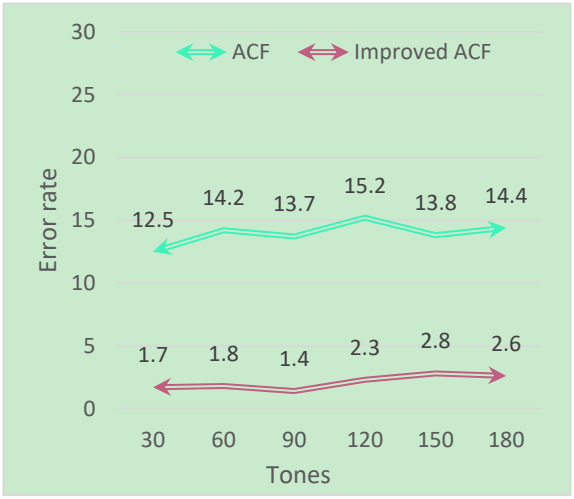
In this paper, 180 tones are selected from MIR-1K dataset and MedleyDB dataset for experiments. The selected dataset is evenly distributed on each attribute, and the samples are

representative. The test results are also universal and have certain reference value.

The error rates of 180 tones extracted by the two algorithms under different data sets are shown in Figure 6.



(a) Error rate of two algorithms extracted from MIR-1K dataset



(b) Error rate of two algorithms extracted from MedleyDB dataset

Figure 6: MIR-1K dataset and MedleyDB dataset

As shown in Figure 6, in Figure 6 (a), it can be seen that the error rate of ACF under MIR-1K dataset is 21.1% when 30 tones need to be extracted. The error rate is 17.5% when 60 tones need to be extracted. The error rate is 18.3% when 90 tones need to be extracted. When 120 tones need to be extracted, the error rate is 17.3%. When 150 tones need to be extracted, the error rate is 17.9%. The improved ACF has an error rate of 2.3% when 30 tones need to be extracted. The error rate is 2.5% when 60 tones need to be extracted. The error rate is 3.6% when 90 tones need to be extracted. The error rate is 5.4% when 120 tones need to be extracted. When 150 tones need to be extracted, the error rate is 5.8%. It can be found that the error rate of ACF is higher than that of improved ACF.

In Figure 6 (b), it can be seen that the error rate of ACF in MedleyDB dataset is 12.5% when 30 tones need to be extracted. The error rate is 14.2% when 60 tones need to be extracted. The error rate is 13.7% when 90 tones need to be extracted. The error rate is 15.2% when 120 tones need to be extracted. When 150 tones need to be extracted, the error rate is 13.8%. The improved ACF has an error rate of 1.7% when 30 tones need to be extracted. The error rate is 1.8% when 60 tones need to be extracted. When 90 tones need to be extracted, the error rate is 1.4%. When 120 tones need to be extracted, the error rate is 2.3%. The error rate is 2.8% when 150 tones need to be extracted.

4.2 Classification Effect and Melody Difference of Different Algorithms

In this paper, 100 pieces of classical music, 100 pieces of hip-hop music and 100 pieces of country music are selected for experimental research on music classification. The classification effects of ACF algorithm and improved ACF algorithm are shown in Table 3.

Table 3: Classification accuracy of different algorithms for different music

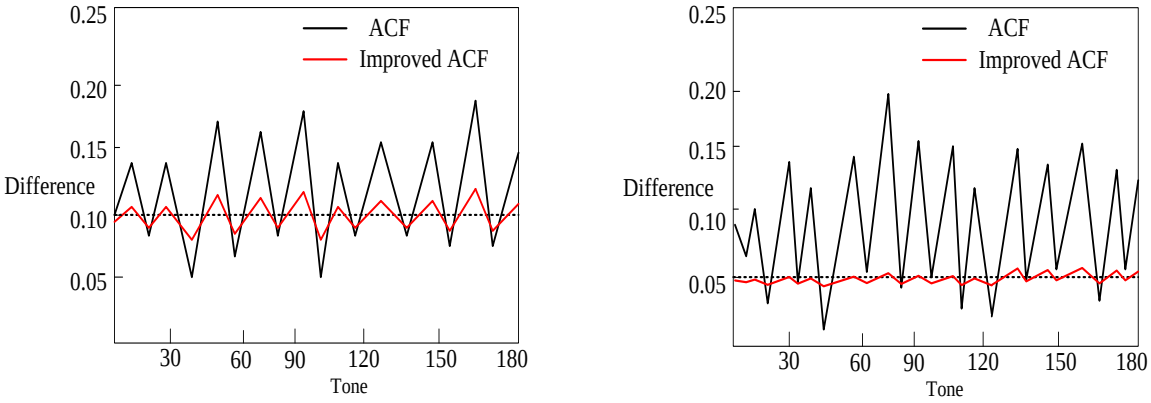
Music classification	ACF	Improved ACF
Classical music	73%	84%

Country music	75%	88%
Hip hop music	78%	94%

As shown in Table 3, the classification accuracy of ACF algorithm for classical music is 73%, for country music is 75%, and for hip-hop music is 78%. The classification accuracy of the improved ACF algorithm is 84% for classical music, 88% for country music and 94% for hip-hop music.

Using ACF algorithm and improved ACF algorithm to classify music, it can be found that the improved ACF algorithm has a high accuracy, and the improved ACF algorithm can reach 94% at most.

Then, this paper compares the melody differences of the two algorithms under different data sets, as shown in Figure 7.



(a) Melody difference between two algorithms in MIR-1K dataset

(b) Melody difference between two algorithms in MedleyDB dataset

Figure 7: Melodic differences between the two algorithms under different data sets

As shown in Figure 7, it can be seen in Figure 7 (a) that the melody difference of ACF algorithm under MIR-1K dataset is much greater than that of the improved ACF algorithm. The melody difference of ACF algorithm fluctuates greatly, while the melody difference of improved ACF algorithm fluctuates slightly and is relatively stable.

Figure 7 (b) shows that the melody difference of the improved ACF algorithm in MedleyDB dataset is much smaller than that in MIR-1K dataset, indicating that the improved ACF algorithm has more advantages.

The improved ACF algorithm has higher compression ratio, more stable repetition times and less melody difference. Therefore, more accurate pitch value estimation can be obtained. It is expected to play an important role in speech analysis, music retrieval, audio classification, audio data visualization and other fields.

4.3 Comparison of Extraction Efficiency of Different Algorithms

In order to verify that the improved ACF algorithm proposed in this chapter is more efficient, the extraction time of 60 pieces of music by ACF algorithm and the improved ACF algorithm is compared, as shown in Table 4.

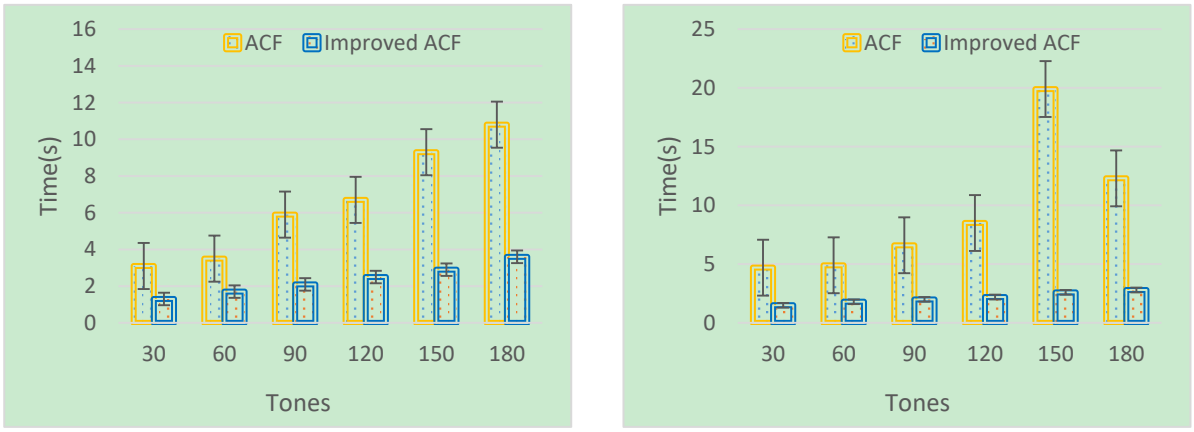
Table 4: ACF algorithm and improved ACF algorithm for music extraction time (s)

Number of music	ACF	Improved ACF

10	1.5	0.5
20	1.7	0.9
30	2.8	1.5
40	3.2	1.7
50	3.9	2.1
60	4.6	2.5

As shown in Table 4, when the number of music to be extracted is 10, the extraction time of ACF algorithm and improved ACF algorithm for music is 1.5s and 0.5s. When the number of music to be extracted is 20, the extraction time of the two algorithms is 1.7s and 0.9s. When the number of music to be extracted is 30, the extraction time of the two algorithms is 2.8s and 1.5s. When the number of music to be extracted is 60, the extraction time of the two algorithms is 4.6s and 2.5s. With the increase of the number of music to be extracted, the extraction time of the two algorithms is also increasing. However, the extraction time of the improved ACF algorithm is always lower than that of the ACF algorithm, which also shows the effectiveness of the improved ACF algorithm for audio extraction.

In order to verify the scientificity of the experiment, the next two algorithms are run separately on two datasets. 180 tones are extracted to record the extraction time of the two algorithms, as shown in Figure 8.



(a) Extraction time of two algorithms in MIR-1K dataset

(b) Extraction time of two algorithms in MedleyDB dataset

Figure 8: Extraction time of tones and melodies by two algorithms under different data sets

As shown in Figure 8, it is observed from Figure 8 (a) that the time taken for ACF algorithm and improved ACF algorithm to extract melody when the tone is 30 under MIR-1K dataset is 3.1s and 1.3s respectively. The two algorithms take 3.5s and 1.7s respectively when the pitch is 60. The two algorithms spend 5.9s and 2.1s respectively when the pitch is 90. The two algorithms take 6.7s and 2.5s respectively when the pitch is 120. When the pitch is 180, the two algorithms spend 10.8s and 3.6s respectively.

From Figure 8 (b), it is observed that the time taken for ACF algorithm and improved ACF algorithm to extract melody when the tone is 30 under MedleyDB dataset is 4.7s and 1.5s respectively. The two algorithms take 4.9 s and 1.8 s respectively when the pitch is 60. The two algorithms take 6.6s and 2s respectively when the pitch is 90. The two algorithms take 8.5 s and 2.2 s respectively when the pitch is 120. The two algorithms take 12.3s and 2.8s respectively when the pitch is 180.

5 Conclusions

With the rapid development of computer technology, the dynamic visualization of vocal music melody has attracted more and more attention, and various dynamic visualization systems of vocal music melody have emerged as the times require. Dynamic visualization of vocal music melody is a new visual technology. It has paid more attention to the connection between sound and image, and is the focus of new research fields such as music, image processing, virtual reality, etc. It has extensive applications in entertainment, education, art, commerce, etc. Data visualization technology is a simple and efficient method, which can solve the problems of low efficiency and high professional threshold in traditional voice analysis. The dynamic visualization of vocal music melody based on audio feature extraction and image processing is aimed at the problems of low efficiency and high professional threshold of traditional vocal music melody dynamic visualization. It can simplify the vocal music analysis process, which helps hearing disabled people perceive vocal music content, and provides more rich audio-visual experience. Vocal music is not only a form of entertainment, but also allows composers and performers to express their emotions. Sound emotional components make dynamic visualization of vocal music melody an important aspect of music.

About the Author

Jia Kuang was born in LouDi, HuNan.P.R. China, in 1980 He received the Doctoral degree from Yunan Minzu University, P.R. China. Now, he works in college of music, Hengyang Normal University. His research interests include folk music culture and sociology.
E-mail: hysfxykj001@126.com

Dihua Zhang was born in Hengyang, Hunan, P.R. China, in 1972. He received the Bachelor's degree from Hengyang Normal University, P.R. China. Now, he works in the College of Music, Hengyang Normal University. His research interests include piano tuning and folk music culture.
Email: 1165849680@qq.com

References

[1] Lima, H. B., Dos Santos, C. G. R., & Meiguins, B. S. (2021). A survey of music visualization techniques. ACM Computing Surveys, 54(7), 1-29.

[2] Arai, K., Hirao, Y., Narumi, T., Nakamura, T., Takamichi, S., & Yoshida, S. (2023). TimToShape: Supporting practice of musical instruments by visualizing timbre with 2D shapes based on crossmodal correspondences. In Proceedings of the 28th International Conference on Intelligent User Interfaces (pp. 850-865).

[3] Graf, M., Opara, H. C., & Barthet, M. (2021). An audio-driven system for real-time music visualisation. In Audio Engineering Society Convention 150.

[4] Park, E. J. (2025). Music dynamics visualization for music practice and education. Multimedia Tools and Applications, 84, 36145-36161.

[5] Erdmann, M., von Berg, M., & Steffens, J. (2025). Development and evaluation of a mixed reality music visualization for a live performance based on music information retrieval. Frontiers in Virtual Reality, 6, 1552321.

[6] Huang, J., Weber, C. J., & Rothe, S. (2025). An AI-driven music visualization system for generating meaningful audio-responsive visuals in real-time. In Proceedings of the 2025 ACM International Conference on Interactive Media Experiences (pp. 258-274).

[7] Yu, X. (2025). The impact of music visualization model by using internet of things techniques and deep neural network. Scientific Reports, 15, 39659.

[8] Wei, H., Cao, X., Dan, T., & Chen, Y. (2023). RMVPE: A robust model for vocal pitch estimation in polyphonic music. In Proceedings of Interspeech 2023 (pp. 5421-5425).

[9] Zhang, W., Yan, L., Zhang, Q., & Gao, J. (2023). Graph modeling for vocal melody extraction from polyphonic music. Applied Acoustics, 211, 109491.

[10] Yu, S., Yu, Y., Sun, X., & Li, W. (2023). A neural harmonic-aware network with gated attentive fusion for singing melody extraction. Neurocomputing, 521, 160-171.

[11] Yu, S., Liu, J., Yu, Y., & Li, W. (2024). A scalable sparse transformer model for singing melody extraction. In 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1071-1075).

[12] Jing, J., Hu, Y., Huang, H., He, L., & Ou, Z. (2025). A joint network for singing melody extraction from polyphonic music with attention aggregation and self-consistency training. In Proceedings of Interspeech 2025 (pp. 3100-3104).

[13] Hsu, C.-L., & Jang, J.-S. R. (2010). On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. IEEE Transactions on Audio, Speech, and Language Processing, 18(2), 310-319.

[14] Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2014). MedleyDB: A multitrack dataset for annotation-intensive MIR research. In Proceedings of the 15th International Society for Music Information Retrieval Conference (pp. 155-160).

[15] Ju, Y., Wu, C. Y., Cortiñas-Lorenzo, B., Yang, J., Deng, J., Fan, F., & Lui, S. (2024). End-to-end automatic singing skill evaluation using cross-attention and data augmentation for solo singing and singing with accompaniment. In Proceedings of the 25th International Society for Music Information Retrieval Conference (pp. 493-500).