Automatic Vocal Melody Extraction Via Quadratic Fluctuation Equation: A Comparative Analysis

Xiaoquan He, Fang Dong*
Arts Academy of Shaoxing University, Shaoxing, China, 312000
*E-mail: hxq@usx.edu.cn
*Corresponding author

Keywords: digitizing, extraction, discrimination, music, vocal melodies, quadratic fluctuation equation (QFE)

Recieved: July 11, 2024

The extraction and recognition of vocal melodies from music data is an intricate but necessary step in digitized music technology. Efficient preprocessing methods are essential for precise musical signal evaluation and processing. Conventional techniques for automating this task include Convolutional Recurrent Neural Network-Conditional Random Field (CRNN-CRF), Non-Harmonic Adaptive Network-Global Average Filtering (NHAN-GAF), and Frequency-Aware Multi-Objective Regression (FA-MOR), but they have constraints like lower accuracy and higher false alarm rates. These techniques frequently fail to sustain high precision in differentiating vocal melodies, resulting in suboptimal efficiency. Objectives: To tackle these drawbacks, the Quadratic Fluctuation Equation (QFE) is proposed as an innovative technique for automatically extracting vocal melodies. Methods: The QFE technique uses a Wiener filter and a penalized procedure to generate a dual-objective metric that efficiently manages phase discrepancies and reduces errors in inversion operations. This technique is especially good at compensating for problems like cyclical jumps in the frequency domain, which are common pitfalls in conventional techniques. Comprehensive computational experiments were carried out on a dataset of 373 ancient Chinese instrumental music pieces. The dataset, which included spectrograms from 17 various instruments, presented a solid foundation for assessing the effectiveness of the Quadratic Fluctuation Equation. Results: Experimental findings show that the Quadratic Fluctuation Equation surpasses previous approaches with an accuracy of 98%, which is an important advancement over modern techniques. The Quadratic Fluctuation Equation also performed well in terms of voice recall value, false alarm ratio, raw pitch precision, and raw chroma level. Conclusion: Overall, the Quadratic Fluctuation Equation technique is a robust solution for extracting and discriminating vocal melodies, with higher accuracy and dependability than previous methods. The findings highlight the capacity of the Quadratic Fluctuation Equation to advance the area of digitized music evaluation and signal processing.

Povzetek:Razvita je enačba kvadratnih fluktuacij (QFE) za ekstrakcijo vokalnih melodij, ki presega tradicionalne metode, zmanjšuje fazne neskladnosti in napake ter izboljšuje digitalno obdelavo glasbenih signalov.

1 Introduction

Melody extraction is the collection of frequency components reflecting the dominant melodic line of a polyphony musical. This is a top goal in the field of music information retrieval MIR, with applications including humming-based inquiries, cover song recognition, and sing voice splitting [1]. This method is extremely challenging for two main reasons: First, in polyphonic music, it is typical for numerous instruments and singing voices to be performed simultaneously and jumbled by the harmonic structure, which makes it difficult to distinguish and identify F0 readings for particular devices. Next, while it's true that F0 values can be accurately detected, it remains laborious to assess if they belong to the leading melody [2]. Users need an appliance that not only meets their natural inclination to listen to and identify music, but also facilitates the discovery of new music, but also improves the feedback and retrieval impact. Extraction of the main melody and estimate of many pitches are major subjects in the field of music information retrieval. Main melody extraction frequently known as "melody extraction", the computer mechanically extracts melodies analyzing the audio content of a piece of music and extracting the song's main melody [3]. The principal melody is the basis of the song and is the foundation for a variety of applications, including song score recognition, pitch analyses, and musical topic assessment. Depending on the number of simultaneous sound sources, music can be categorized as either monophonic or polyphonic from the standpoint of music signal processing [4]. Singing melody extraction has been a hot topic in the music information retrieval field because of its many downstream applications, including music retrieval, cover song recognition, and music transcribing. The harmonic components of the polyphonic audio have a complicated pattern [5]. A possible method for identifying cover versions is to isolate the primary melody, and accompaniment, and search for both components separately. In addition to its usage in identifying cover versions, primary melody extraction has a wide range of other musical applications. Several basic elements and associated harmonics compose a polyphonic music signal [6]. It extracts the meanings and qualities of objects directly from voice data and then searches a vast library of voice data for voice data with similar attributes. Melody-based music recovery includes humming-based song recovery as a subcategory [7]. Fundamental frequency (f0) is defined as the rate of vocal fold vibrations during song or speech production. Although it is true that while both speech and music are created by comparable vocal assembling, they are radically different in terms of both production and perception. Although discourse provides a language-driven message, music transmits both songs and verses. Moreover, the effect of source-channel link miracles is more notable in music than in speech, according to vocalists who produce regulated variations in f0 by rapidly adjusting the posture of the throat in response to perceptual inputs. Therefore, speakers are often less concerned about the variations in f0. Moreover, the f0 zone is larger and the individual sound units last longer in a melody than in speech [8].

A Joint detection and categorization (JDC) system that concurrently detects singing voices and calculates pitch, the JDC system is made up of a primary system that forecasts the pitch curves of the vocal melody and a supporting system that aids in voice identification. Constructed using a convoluted recurrent neural network with remnant links, the main network involves predicting pitch labels that cover the vocal range, in addition to non-voice status [9]. A unique approach for separating singing voices combines the benefits of the original extract method with the deep neural network (DNN). They utilize DNN's excellent feature extraction capability to retrieve the fundamental frequency of singing music, and then they use the non-negative Network Equations method and the normal vocal resonant concept to create smooth masking for the finished isolated audio [10]. Automated text-to-music harmonization lines up the rhymes with the combined singing audio (performing voice with music playing). An automated speech detection algorithm can accomplish this synchronization [11]. The goal of this study is to evaluate the effectiveness of guided percussive vocal training vs linguistic treatment in improving communicative effectiveness [12]. Vocal treatment, often provided by a Speech Language Pathologist (SLP-V) via telemedicine, is indeed the usual. Despite this "black box" representation, there are several well-established medical therapies that have commonly prescribed objectives and clinical aims yet show signs of subpar performance. Several European singers and voice coaches use the Comprehensive Voice Training (CVT) [13]. "Global Voice Prevention and Treatment Model (GVPTM)" therapeutic conditions have been evaluated [14] using voice health academic instructors in both in-person and telepractice settings. Throughout the presentations, surface electromyography (sEMG) was utilized to assess the muscle strength associated with breathing and position. Position differences in phonatory muscles sEMG activation and aerodynamics voice characteristics have been examined utilizing the multivariate Kruskal-Wallis test[15]. One way to evaluate the quality of a musician's performance is via the application of a spatial detection method utilizing sensor data. Based on the unique qualities of each vocal line, they use the adaptable cascading retrieving control system to track down and extract their achievements. Mined voice audio streams are attributed using a sensor's spatially localization technique and a large database of high-quality vocals. To match the extracted voice with the writing, a typical method uses Dynamic Time stretching to maximize mutual information. To find the best text-to-voice orientation, a new method called Connectionist Temporal Classification (CTC) has been presented. Even though this method yields remarkable results, it can only be used with very short files since the storage cost of the optimal orientation searching maintains a linear input length [17]. The evaluation of vocal music instruction for performers is affected by several things. The score results of evaluators are heavily impacted by subjective considerations. The Back Propagation Neural Network (BPNN) is a revolutionary technique that can replicate any nonlinear continuous function to a given degree of precision. The Back Propagation Neural Network is part of the widely used adaptive feedforward learning network [18].

2 Related works

In the area of vocal melody extraction and voice training, different approaches have been created to address issues such as precision, alignment, and performance assessment. Conventional methods, like convolutional neural networks and acoustic models, have presented useful knowledge, but they have constraints, like low accuracy in noisy settings and difficulty managing polyphonic audio. This section provides a review of key contributions from associated works, concentrating on their objectives, methodologies, findings, and constraints to present an in-depth knowledge of progress and existing gaps in this area.

Table 1: Summary of related work in vocal melody extraction and voice training

Reference No	Objective	Methodology	Result	Limitations
[9] Kum & Nam, 2019	Joint identification and classification of singing voice melody.	A convolutional recurrent neural network (CRNN) method with remaining links for pitch forecasting and voice classification.	Enhanced effectiveness through efficient identification of singing voice and pitch.	Low precision in noisy settings and difficulties with overlapping voices.
[10] Durrieu et al., 2010	Unsupervised main melody extraction from polyphonic audio signals	Deep neural networks are used in the source/filter model to distinguish between vocal melody and accompaniment.	Precise extraction of the singing voice's basic frequency.	Constrained resilience in complicated polyphonic settings; phase discrepancies cause errors.
[11] Sharma et al., 2019	Automated lyrics-to-audio alignment for polyphonic music.	Singing-adapted acoustic models employing an automatic speech identification algorithm.	Lyrics are now better aligned with singing audio.	Accuracy problems in noisy settings and challenges in managing various audio conditions.
[12] Jungblut et al., 2022	Assess the effectiveness of directed rhythmic-melodic voice training for treating chronic non-fluent aphasia.	Behavioral and imaging outcomes of directed rhythmic-melodic voice training.	Significant enhancement in communicative capabilities of patients with non-fluent aphasia.	Constrained applicability to other kinds of voice disorders or nonfluent speech conditions.
[13] McGlashan et al., 2022	Assess the effectiveness of the Complete Vocal Technique (CVT) in patients with muscle tension dysphonia using telehealth.	Telehealth-delivered CVT for enhancing voice and function	Enhanced vocal function and excellence in patients	Real-time voice coaching is limited by telehealth and a small sample size.
[14] Grillo, 2021	Evaluate the Global Voice Prevention and Therapy Model for student teachers through in-person and telepractice Estill Voice Training.	The VoiceEvalU8 app evaluates voice quality in student teachers in both in-person and telepractice situations.	Enhanced voice health and function in student teachers.	There is constrained follow-up data on long-term voice health enhancements.
[15] Castillo-Allendes et al., 2022	Investigate muscle activity and aerodynamic voice modifications at	Surface electromyography (sEMG) is used to assess phonatory	Discovered posture-related modifications in voice muscle	A pilot study with a small sample size and the requirement

	various body postures.	muscle activity across various postures.	activity and aerodynamics.	for more various postures.
[16] Hongtao & Li, 2022	Assess the accuracy of vocal art efficiency using sensor space localization.	Monitor vocal performance using a sensor space localization technique.	Correct assessment of vocal efficiency in terms of spatial positioning.	Constrained adaptability to multiple vocal genres and efficiency settings.
[17] Doras et al., 2023	Text-to-voice alignment for lengthy audio recordings.	Linear Memory Connectionist Temporal Classification (CTC) for text-to-speech alignment	Attained precise text-to-voice alignment for lengthy recordings.	Storage expenses and constraints in processing extremely long files
[18] Cao, 2022	Assess vocal music teaching utilizing Backpropagation Neural Network (BPNN).	BPNN evaluation of nonlinear functions for vocal music instruction.	Enhanced assessment of vocal music instruction	Subjective bias in evaluator scoring influences the findings.

The studies reviewed demonstrate significant advances in vocal melody extraction, alignment, and voice training, with multiple methods improving efficiency in particular circumstances. However, typical difficulties, such as managing complicated polyphonic audio, noise sensitivity, and subjective biases in assessment, persist across these techniques. These constraints highlight the ongoing necessity for more flexible and resilient solutions to these problems. The presented QFE technique closes these gaps by establishing a more efficient way to handle phase discrepancies, decrease false alarms, and enhance the overall accuracy of vocal melody extraction.

3 Materials and method

As previously stated, the main objective of our research is the extraction and discrimination of musical signals. The procedure of extracting music characteristics is the major emphasis of this part. For this study, the Chinese dataset was utilized. Preprocessing is done by using a pre-emphasis filter. The research flow is depicted in Figure 1.

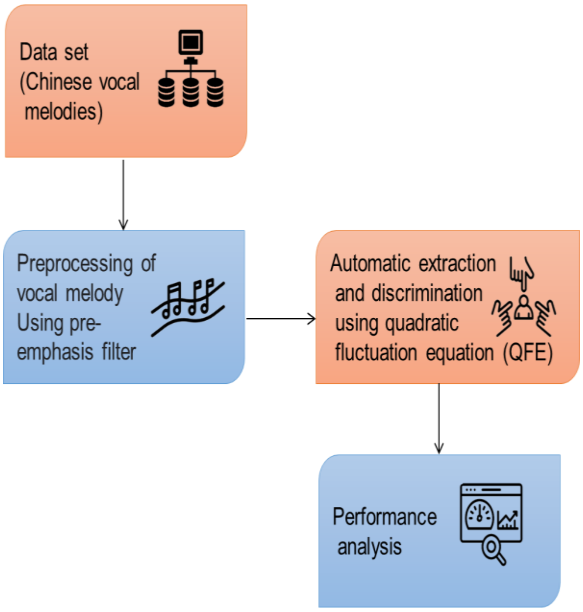
Data set (Chinese vocal melodies)
Preprocessing of vocal melody Using pre-emphasis filter
Automatic extraction and discrimination using quadratic fluctuation equation (QFE)
Performance analysis

Figure 1: Research flow

3.1 Data set

A comprehensive dataset was created, comprising 373 instrumental music pieces in spectrogram form. These pieces showcase 17 various kinds of instrumentation, executed by over 60 musicians. The main objective of creating this dataset was to compile a large collection of ancient Chinese music to study vocal melody extraction. The music was primarily sourced from professional compact discs (CDs), which were selected for their superior recording excellence and efficiency. Professional CDs were chosen over other sources of music audio for multiple factors: they frequently show better sound quality as a result of modern recording techniques utilized in commercial studios, and getting such a huge volume of excellent data on their own would be prohibitively costly. The Short-Time Fourier Transform was used to convert each music piece in the dataset into a spectrogram, which is a visual representation of frequencies over time. This method guarantees that each audio sample is precisely represented in terms of its frequency elements, which are critical for the subsequent evaluation [19].

3.2 Preprocessing of vocal melody using pre-emphasis filter

A pre-emphasis filter was used during the preprocessing of vocal melodies, which is an important step in enhancing signal quality. This filter tackles the problem where higher frequency elements of an audio signal appear less prominent than lower frequency elements. A pre-emphasis filter boosts the magnitudes of higher frequencies, thereby balancing the frequency range and improving the total signal-to-noise ratio (SNR). This procedure aids in the resolution of numerical problems that may arise during the computation of the Fourier transform. Particularly, the pre-emphasis filter compensates for a 6 dB/octave reduction in signal strength above 8 kHz, preserving high-frequency details. Furthermore, the pre-emphasis step aids in the elimination of any DC-level shifts that may influence the accuracy of the signal representation. The audio signal is divided into frames utilizing the window technique, with each frame ranging from 33 to 100 frames per second. The frame length and displacement are adjusted to guarantee seamless changes between successive frames, with the displacement typically set to one-third of the frame length. This careful preprocessing guarantees that the audio signal's features are precisely recorded and evaluated for future steps in the study.

Concerning reproducibility, the dataset is not presently accessible to the public. Further attempts will be made to present accessibility to the dataset to assist reproducibility and future study in this field.

3.3 Automatic extraction and discrimination using quadratic fluctuation equation (QFE)

The music signal concept is made up of three functions: glottal activation function, vocal cords modulating function, and mouth radiated function. These functions are based on the features of the vocal cords modeling of the music signal. Equation 1 illustrates how these three functions are connected in sequence to create the music signal creation mechanism.

$$T(z) = G(z)V(z)M(z) \quad (1)$$

Lossless sound channels and formant simulations are popular representations of the vocal cords. The vocal cords oscillation, which happens in specific frequency ranges, has an impact on the activation wave of the audio input. The resonance's highest point is the maximum created by the contour of the spectral curve at the resonance wavelength. The all-pole concept of the vocal cords represents generic vowels, whereas the zero-pole form represents non-general vowels and the majority of vowels. Equation 2 defines the transfer characteristic formula of a second-order resonance.

$$F_i(y) = \frac{z_i}{1-a_i z^{-1}-b_i y^{-2}} \quad (2)$$

The formant structure of the audio stream is generated by obtaining several F_i linear arrangements.

$$F(y) = \lim_{N \to \infty} \prod_{i=0}^{M-1}[(1 - z_i) \times (1 - a_i y^{-1} + b_i y^{-2})]$$
(3)

We refer to the proportion of the sound signal to the voice tract's yield pulse speed as the radioactive amplitude, neglecting the fact that the open field of the mouth is significantly narrower than the face total area, and we deduce the radioactive resistive interpretation in equation 4 because the resonant framework of the sound signal is an articulation in the type of all poles.

$$y_L(\Omega) = jM_r Q_r \Omega \times (Q_r - jM_r \Omega)^{-1} \quad (4)$$

Equation 5 specifies the goal value of the QFE inverting in the time dimension.

$$B(n) = 0.5(\Delta w)^2 = 0.5(w - p)^2 \quad (5)$$

The pulse field for the current experiment is represented by p, the waveform field by w, and the residue by ∂w. The QFE inversion's residue formula is found in equation 6.

$$\Delta w_i = Sup\{p_i - w_i\} \quad (6)$$

A cyclical leap will happen at this point if the pitch discrepancy between the projected information and the actual data is more than half a loop. Since the original sample is often inaccurate when applied to real core samples, it is susceptible to loop hopping, which has a significant influence on the inversion. Depending on this, we suggested adding a penalty phrase to the goal function to limit it and prevent the cyclical leap.

The QFE inversion is suggested as a way to reduce the impact of cyclical leaps on the reversal. It may be reversed with an inadequate beginning structure and yet provide outcomes that are close to perfect. The QFE inversion technique and theory are distinct from the conventional full-wave equation inverted approach. Rather than employing straight subtraction in this case, the filter and one of the data sources are convolutional before being utilized to deduct from the other given dataset. The incidence of cyclical jumps may be effectively suppressed by the adaptive full-wave formula inversion.

A signal's combination with its impact signal, f(t), yields ft itself. The waveform field w is produced by the convolution of the stress value with the waveform field quantity d. u.d may be produced when the projected waveform field information and the actual waveform field information are highly similar. The modeled information is mixed with the estimated filter factors once the filter parameters have been computed. The phase gap between the modeled and actual data is steadily decreased by continuous repetition, and the cycle leap is effectively controlled. At the same time, the predicted values continue to become closer to the genuine data. The filter factor resembles or progressively transforms into a stress function. At this point, the discrepancy between the actual data and the modeled information is as small as possible, and a perfect inversion effect is ultimately realized. Upward QFE inversion is the name of this technique. The difference between the actual and modeled data may also be narrowed by repetition when the actual data is mixed with the filter factors and contrasted with them. This technique is known as the eventual responsive QFE inversion. By using these vocals main melodies are extracted and discriminated. Algorithm 1 demonstrates the suggested Quadratic Fluctuation Equation (QFE) algorithm.

Algorithm 1: Quadratic Fluctuation Equation (QFE) Algorithm

Input	**:**	Spectrogram of Music: Audio frequency time representation. Pre-processed Signal: Audio signal processed with a pre-emphasis filter. Filter Parameters: Values utilized to equalize frequency and optimize SNR.
Output	**:**	Extracted Vocal Melody: Isolated melody from the music. Modeled Data: Predicted melody values. Inversion Results: Adjusted the result of the QFE inversion procedure.
Step 1	**:**	Prepare Data: Transform audio to spectrogram and use pre-emphasis filtering.
Step 2	**:**	Model Functions: Define glottal activation, vocal cord modulation, and mouth-radiated functions.
Step 3	**:**	Simulate Vocal Cords: Utilize formant simulations to model vocal resonance.
Step 4	**:**	Generate Formant Structure: Obtain linear configurations for the formant structure.
Step 5	**:**	Compute Amplitude: Calculate the ratio between sound signal and vocal tract pulse speed.
Step 6	**:**	Apply QFE: Utilize QFE to model and invert the vocal signal, and manage cyclical jumps with a penalty term.
Step 7	**:**	Improve Findings: Iterate to decrease the discrepancies between modeled and actual data.
Step 8	**:**	Extract Melody: Complete vocal melody extraction by reducing variances between real and modeled data.

The Quadratic Fluctuation Equation (QFE) algorithm separates vocal melodies from instrumental music by initially converting the audio to a frequency-time representation known as a spectrogram and then using a pre-emphasis filter to improve high-frequency elements. It then models the vocal signals with functions that simulate vocal cord vibrations and formant structures. The algorithm manages cyclical errors by computing amplitude and performing the QFE inversion technique to the data. Through iterative refinement, the QFE algorithm reduces discrepancies between forecasted and actual data, leading to precise extraction of the vocal melody from the music.

4 Results

This section demonstrates the evaluation of the quadratic fluctuation equation in the process of extraction of vocal main melodies. The performance metrics used for evaluation include accuracy, voice recall value, false alarm ratio, raw pitch precision, and raw chroma level. The existing techniques used for comparison are Convolutional Recurrent Neural Network-Conditional Random Field (CRNN-CRF) [20], neural harmonic-aware network with gated attentive fusion (NHAN-GAF) [21], and Frequency amplitude and multi-octave relation (FA-MOR) [22].

4.1 Accuracy

Accuracy is defined as the percentage of frames in the extraction when melodies and voices are accurately predicted. The accuracy of the mechanism may be calculated as a measure of its efficacy. Figure 2 shows the accuracy of the extraction and discrimination of the vocal main melodies of the proposed and existing methods. Table 2, shows the accuracy outcomes. This demonstrates that the QFE offers more accurate melody extraction than the standard methods.

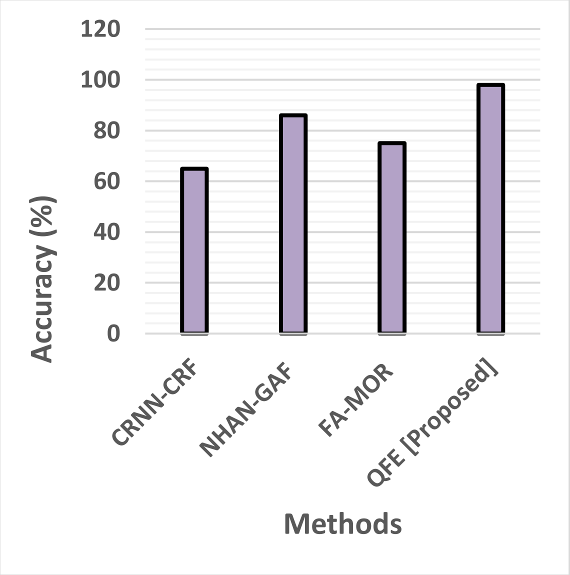

Figure 2: Accuracy of the extraction of vocal melodies

Table 2: Outcomes of accuracy

Methods	Accuracy (%)
CRNN-CRF	65
NHAN-GAF	86
FA-MOR	75
QFE [Proposed]	98

4.2 Voice recall value

The voiced frames with accurate voicing estimation from the voiced frames are referred to as recall. Recall is the capacity of a system to identify all the pertinent answers inside a given dataset. Figure 3 shows the voice recall value of the extraction and discrimination of the vocal main melodies of the proposed and existing methods. Table 3 shows the voice recall value outcomes. This shows that the QFE is capable of providing high recall value.

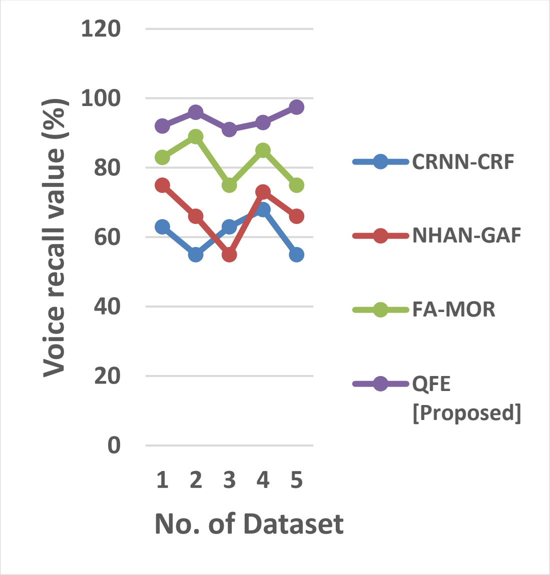

Figure 3: Voice recall value of the extraction of vocal melodies

Table 3: Outcomes of voice recall value

No. of Dataset	Voice recall value (%)			
	CRNN-CRF	NHAN-GAF	FA-MOR	QFE [Proposed]
1	63	75	83	92

2	55	66	89	96
3	63	55	75	91
4	68	73	85	93
5	55	66	75	97.5

4.3 False alarm ratio

Unvoiced sequences when unvoicing is incorrectly calculated from the unvoiced sequences are known as false alarm ratios. It indicates that the method's incorrect extraction of the vocal melody. It demonstrates that the QFE has a low false alarm ratio, which results in fewer extraction errors. Figure 4 shows the false alarm ratio of the extraction and discrimination of the vocal main melodies of the proposed and existing methods. Table 4 shows the false alarm ratio outcomes.

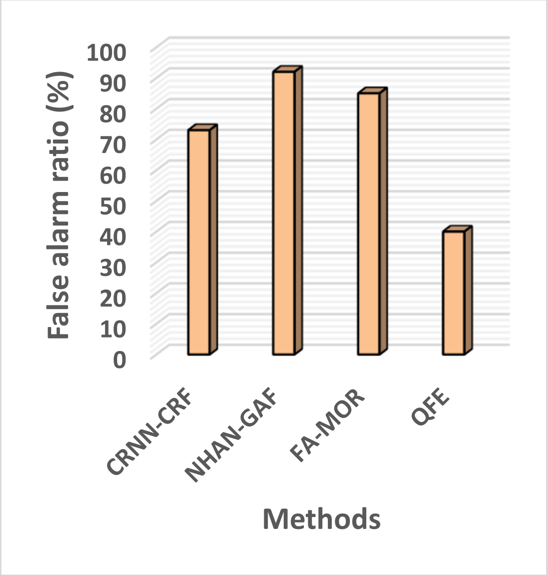

Figure 4: False alarm ratio of the extraction of vocal melodies

Table 4: Outcomes of false alarm ratio

Methods	False alarm ratio (%)
CRNN-CRF	73
NHAN-GAF	92
FA-MOR	85
QFE [Proposed]	40

4.4 Raw pitch precision

The voiced frames with accurate pitch estimation from the voiced sequences are referred to as raw pitch precision. The accuracy of vocal melody predictions is measured by precision. It demonstrates the better raw pitch precision of the QFE and demonstrates its reliability in extraction. Figure 5 shows the raw pitch precision of the extraction and discrimination of the vocal main melodies of the proposed and existing methods. Table 5 shows the raw pitch precision outcomes.

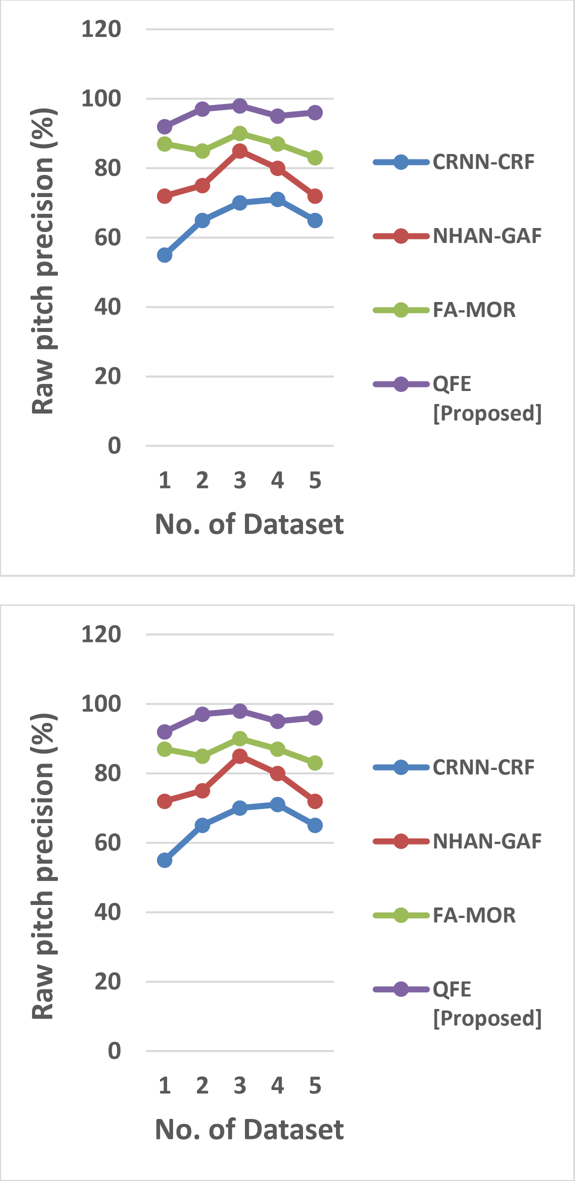

Figure 5: Raw pitch precision of the extraction of vocal melodies

Table 5: Outcomes of raw pitch precision

No. of Dataset	**Raw pitch precision (%)**			
	CRNN-CRF	NHAN-GAF	FA-MOR	QFE [Proposed]
1	55	72	87	92
2	65	75	85	97
3	70	85	90	98
4	71	80	87	95
5	65	72	83	96

4.5 Raw chroma level

The voiced frames when chromas are accurately approximated from the voiced frames are referred to as the raw chroma level. Figure 6 shows the raw chroma level of the extraction and discrimination of the vocal main melodies of the proposed and existing methods. Table 6 shows the raw chroma level outcomes. It demonstrates that the QFE Raw has a greater chroma level than the other techniques and demonstrates the dependability of its extraction.

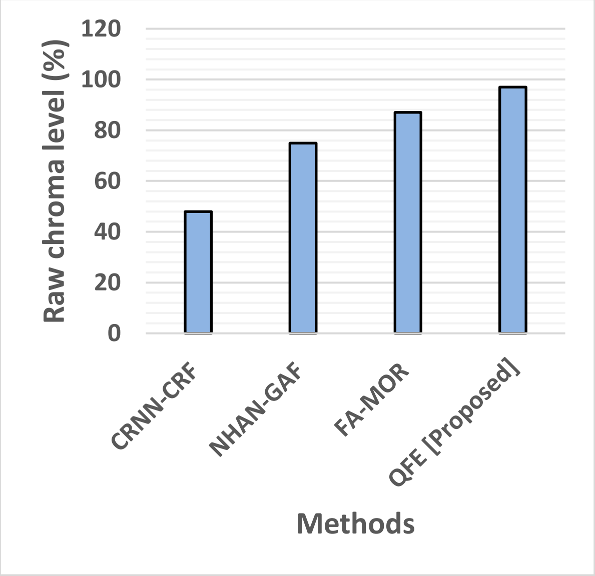

Figure 6: Raw chroma level of the extraction of vocal melodies

Table 6: Outcomes of raw chroma level

Methods	**Raw chroma level (%)**
CRNN-CRF	48
NHAN-GAF	75
FA-MOR	87
QFE [Proposed]	97

5 Discussion

The presented Quadratic Fluctuation Equation (QFE) for automated vocal melody extraction outperforms conventional techniques, as shown in Tables 1–5. When comparing metrics such as accuracy, voice recall value, false alarm ratio, raw pitch precision, and raw chroma level, QFE regularly outperforms other techniques. A deeper analysis of these findings presents knowledge of the causes for QFE's outstanding efficiency and shows potential causes of discrepancies.

In terms of accuracy, QFE attains an outstanding 98%, outperforming the CRNN-CRF, NHAN-GAF, and FA-MOR models, which achieve 65%, 86%, and 75%, respectively. This rise is mainly due to QFE's novel incorporation of a penalized method with the Wiener filter, that efficiently eliminates phase discrepancies and cyclical jumps during the inversion procedure. In contrast, conventional approaches such as CRNN-CRF and NHAN-GAF lack the precision required to manage frequency-domain variations, resulting in lower accuracy rates. Sophisticated preprocessing improves QFE's signal-to-noise ratio, enabling more precise melody extraction.

The voice recall value also demonstrates QFE's superiority, with scores ranging from 91% to 97.5%, as opposed to the fluctuating and typically lower recall values of CRNN-CRF (55%-68%), NHAN-GAF (55%-75%), and FA-MOR (75%-89%). QFE's capability to capture a higher proportion of true vocal melodies could be attributed to its quadratic fluctuation method, which excels at handling complicated frequency modulations commonly found in vocal music. Conventional techniques struggle to recognize these finer details, resulting in lower recall values. The structured dataset utilized in the QFE model helps to enhance efficiency by efficiently capturing the complexities of audio signals.

Another important metric where QFE surpasses its competitors is the false alarm ratio, which is 40% lower than the higher rates of CRNN-CRF (73%), NHAN-GAF (92%), and FA-MOR (85%). QFE's sophisticated penalized method is likely to decrease overfitting and the identification of unnecessary or inaccurate melodies, which is an ongoing problem in conventional models.

Competing models' higher false alarm ratios may be attributed to their dependence on easier signal discrimination techniques, which are less efficient at managing noisy or overlapping frequency bands, particularly in datasets with complicated vocal and instrumental interactions.

In terms of raw pitch precision, QFE maintains to lead with values ranging from 92% to 98%, while the other models have substantially lower precision rates: CRNN-CRF (55%-71%), NHAN-GAF (72%-85%), and FA-MOR (83%-90%). The quadratic fluctuation method in QFE is critical in precisely capturing pitch deviations, resulting in improved efficiency by decreasing cyclical jumps and pitch errors. Conventional models, on the other hand, are absent from this level of advance and frequently struggle to sustain high precision in pitch recognition, particularly on difficult datasets. The pre-emphasis filtering used during QFE preprocessing improves high-frequency resolution, which is essential for precise pitch estimation.

Similarly, QFE surpasses other models in terms of raw chroma level, scoring 97%, whilst CRNN-CRF, NHAN-GAF, and FA-MOR fall behind at 48%, 75%, and 87%, respectively. This suggests that QFE is better at capturing the harmonic structure of an audio signal, owing to its higher time-frequency resolution and accurate phase correction methods. Conventional techniques fall short in this regard, especially since they do not provide an identical level of detail when correcting and maintaining harmonic content. The QFE model's spectrogram-based dataset contributes to the higher raw chroma level score by offering a more precise representation of harmonic content.

The observed differences between QFE and conventional models could be attributed to multiple factors. The dataset utilized in this study, which included a wide range of vocal and instrumental elements, was likely too complicated for simpler models such as CRNN-CRF and NHAN-GAF. Their architectures are not intended to handle complex frequency modulations as well as QFE. Furthermore, QFE's architecture, which combines a penalized method with sophisticated time-frequency evaluation, outperforms other methods in managing phase discrepancies and cyclical jumps that are common in complicated vocal melody extraction tasks. QFE's preprocessing methods, especially the incorporation of a pre-emphasis filter, help its greater efficiency metrics by improving the clarity of high-frequency elements and ensuring precise melody and pitch extraction.

Deep learning-based techniques, such as Convolutional Neural Networks (CNNs), have grown in popularity in Music Information Retrieval (MIR) research because of their ability to learn features from complicated data autonomously. In comparison to CNNs, the QFE technique offers numerous computational benefits. Initially, the QFE algorithm is less computationally intensive, needing fewer resources and shorter training times because it concentrates on signal processing and harmonic feature extraction rather than deep learning architectures, which require extensive data and numerous stages of training. Furthermore, QFE has greater generalizability because it relies less on large, labeled datasets, which are frequently needed for CNNs to prevent overfitting. This renders QFE especially useful for applications in which data collection is expensive or limited. While CNNs excel at learning complicated patterns from massive data sets, QFE's lightweight and effective design may be advantageous in situations requiring quick, dependable findings with low computational overhead.

In conclusion, the QFE model for automatic vocal melody extraction outperforms conventional approaches in all important metrics. Because of its sophisticated architecture, resilient dataset managing, and advanced preprocessing methods, it achieves higher accuracy, better voice recall, fewer false alarms, and better pitch and chroma detection. These findings show QFE's ability to substantially improve the precision and dependability of vocal melody extraction, particularly in difficult musical settings where conventional techniques fall short. Other models' efficiency differences can be attributed largely to their easier architectures and the absence of detailed signal processing methods, which QFE efficiently addresses.

6 Conclusion

Predicting the basic harmonic or pitch associated with the music's origin is known as melody extraction. The vocal performance often serves as the primary rhythmic basis in modern music, making it the most frequent responsibility in melodic lines to determine the singing voice's tone. The research on music is more relevant to people's lives since it is a powerful means of expressing and transmitting feelings. While individuals are formed with the capacity to enjoy and recognize music, it is incredibly challenging for machines to evaluate, comprehend, and extract music material. Thus, the study of music data extraction has been greatly addressed by scientific circles. As a result, the QFE inverted approach is the subject of the study in this work. Chinese vocal is used as the database. We defined inversion and described the full-wave formula inversion theory. The goal functional and slope calculation equations are deduced from the inverting of the full-wave solution in the time dimension. By employing this music information is retrieved. Several criteria, including accuracy, voice recall value, false alarm ratio, raw pitch precision, and raw chroma level, were used to assess the QFE's performance. These factors were contrasted with standard techniques. The findings demonstrate that the QFE is more effective at extracting and discriminating vocal main melody. To further increase the functionality of the method, we will examine the challenge of voice improvement in the future.

Reference

[1] Gao, Y., Zhu, B., Li, W., Li, K., Wu, Y., & Huang, F. (2019, May). Vocal melody extraction via DNN-based pitch estimation and salience-based pitch refinement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 1000-1004). IEEE. https://doi.org/10.1109/icassp.2019.8683608

[2] Gao, Y., Zhang, X., & Li, W. (2021). Vocal melody extraction via hrnet-based singing voice separation and encoder-decoder-based f0 estimation. *Electronics,* 10(3), 298. https://doi.org/10.3390/electronics10030298

[3] Liang, S., & Shu, R. (2022). Extraction of Music Main Melody and Multi-Pitch Estimation Method Based on Support Vector Machine in Big Data Environment. *Journal of Environmental and Public Health,* 2022(1), 1074174. https://doi.org/10.1155/2022/1074174

[4] Li, C., Liang, Y., Li, H., & Tian, L. (2021). Main melody extraction from polyphonic music based on frequency amplitude and multi-octave relation. *Computers & Electrical Engineering*, 90, 106985. https://doi.org/10.1016/j.compeleceng.2021.106985

[5] 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. https://doi.org/10.1016/j.neucom.2022.11.086

[6] Kumar, N., Kumar, R., Murmu, G., & Sethy, P. K. (2021). Extraction of melody from polyphonic music using modified morlet wavelet. *Microprocessors and Microsystems*, 80, 103612. https://doi.org/10.1016/j.micpro.2020.103612

[7] Zhang, J. (2022). Music Data Feature Analysis and Extraction Algorithm Based on Music Melody Contour. *Mobile Information Systems*, 2022(1), 8030569. https://doi.org/10.1155/2022/8030569

[8] Loheswaran, K., Subba Ramaiah, V., Srinivasa Rao, S., Malathi, P., Prabu, M., & Niveditha, V. R. (2022). Powerful basic frequency extraction from monophonic signs utilizing versatile sub-band separating. *International Journal of Speech Technology*, 1-14. https://doi.org/10.1007/s10772-021-09874-4

[9] Kum, S., & Nam, J. (2019). Joint detection and classification of singing voice melody using convolutional recurrent neural networks. *Applied Sciences,* 9(7), 1324. https://doi.org/10.3390/app9071324

[10] Durrieu, J. L., Richard, G., David, B., & Févotte, C. (2010). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE transactions on audio, speech, and language processing,* 18(3), 564-575. https://doi.org/10.1109/tasl.2010.2041114

[11] Sharma, B., Gupta, C., Li, H., & Wang, Y. (2019, May). Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models. In ICASSP 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 396-400). IEEE. https://doi.org/10.1109/icassp.2019.8682582

[12] Jungblut, M., Mais, C., Binkofski, F. C., & Schüppen, A. (2022). The efficacy of a directed rhythmic-melodic voice training in the treatment of chronic non-fluent aphasia—Behavioral and imaging results. *Journal of Neurology,* 269(9), 5070-5084. https://doi.org/10.1007/s00415-022-11163-2

[13] McGlashan, J., Aaen, M., White, A., & Sadolin, C. (2022). A Proof-of-Concept Study of The Complete Vocal Technique (CVT), a pedagogic technique used for Performers, in Improving the Voice and Vocal Function in Patients with Muscle Tension Dysphonia (CVT4MDT) delivered by Telehealth: Study Protocol. https://doi.org/10.1186/s40814-023-01317-y

[14] Grillo, E. U. (2021). A nonrandomized trial for student teachers of an in-person and telepractice Global Voice Prevention and Therapy Model with Estill Voice Training assessed by the VoiceEvalU8 app. *American Journal of Speech-Language Pathology*, 30(2), 566-583. https://doi.org/10.1044/2020_ajslp-20-00200

[15] Castillo-Allendes, A., Delgado-Bravo, M., Ponce, A. R., & Hunter, E. J. (2022). Muscle activity and aerodynamic voice change at different body postures: a pilot study. *Journal of Voice. https://doi.org/10.1016/j.jvoice.2022.09.024*

[16] Hongtao, W., & Li, G. (2022). A Method for Evaluating the Accuracy of Vocal Art Performance of Singers' Voices Based on Sensor Space Localization Algorithm. *Mobile Information Systems*, 2022(1), 5248639. https://doi.org/10.1155/2022/5248639

[17] Doras, G., Teytaut, Y., & Roebel, A. (2023). A linear memory CTC-based algorithm for text-to-voice alignment of very long audio recordings. *Applied Sciences,* 13(3), 1854. https://doi.org/10.3390/app13031854

[18] Cao, W. (2022). Evaluating the vocal music teaching using backpropagation neural network. *Mobile Information Systems,* 2022(1), 3843726. https://doi.org/10.1155/2022/3843726

[19] Shen, J., Wang, R., & Shen, H. W. (2020). Visual exploration of latent space for traditional Chinese music. *Visual Informatics*, 4(2), 99-108. https://doi.org/10.1016/j.visinf.2020.04.003

[20] Li, H., Tian, L., & Li, C. (2023). Multi-task melody extraction using feature optimization and CRNN-CRF. *Computers and Electrical Engineering,* 107, 108605. https://doi.org/10.1016/j.compeleceng.2023.108605

[21] 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. https://doi.org/10.1016/j.neucom.2022.11.086

[22] Li, C., Liang, Y., Li, H., & Tian, L. (2021). Main melody extraction from polyphonic music based on frequency amplitude and multi-octave relation. *Computers & Electrical Engineering*, 90, 106985. https://doi.org/10.1016/j.compeleceng.2021.106985