Journal of Intelligent Computing

DLINE JOURNALS

Print ISSN: 0976-9005
Online ISSN: 0976-9013

JIC 2025: 16 (4)
https://doi.org/10.6025/jic/2025/16/4/156-162

Lyrics Analysis from Polyphonic Music Using Sparse Auto encoders and Pitch Salience Analysis

Xiaoyu Zhang
Henan University of Engineering
Zhengzhou, Henan, 450000. China
zhangxiaoyu@haue.edu.cn

ABSTRACT

This work presents a study on extracting vocal melodies from polyphonic music using advanced signal processing and deep learning techniques. It emphasizes the importance of accurately identifying the main melody typically carried by the human voice for applications such as music information retrieval, cover song identification, and copyright protection. The proposed method involves several stages: signal preprocessing (including down sampling, normalization, and Short Time Fourier Transform), note segmentation using the DIS algorithm, and multiple fundamental frequency (F0) estimation within the 70–1000 Hz range. A key innovation is the use of a sparse auto encoder neural network (SAENN) combined with a softmax classifier to distinguish vocal melody from instrumental accompaniment. Trained and tested on the MIR-1K dataset, the model achieves over 85.1% recognition accuracy. Experimental results show that the approach reduces melody localization errors and decreases average extraction time by 0.12 seconds compared to traditional methods. The study also introduces the Average Extraction Time (AET) as a new metric for evaluating computational efficiency. Overall, integrating improved pitch salience computation with deep learning significantly enhances both accuracy and processing speed in melody extraction, offering a robust solution for real-world audio analysis tasks. The findings suggest promising directions for future work in music signal processing and vocal separation technologies.

Keywords: Vocal melody extraction, polyphonic music, Fundamental frequency (F0) estimation, Pitch salience, Sparse auto encoder neural network (SAENN), Melody discrimination, MIR-1K dataset, Average Extraction Time (AET)

Received: 17 August 2025, Revised 7 October 2025, Accepted 19 October 2025

1. Introduction

With the advancement of technology, we are transitioning into a completely digital age [1]. In this scenario, the volume of multimedia content is rapidly escalating, and the effectiveness of traditional manual input Consequently, the algorithm is designed to eliminate these pitches by diminishing the pitch salience function.

techniques for music searching clearly falls short of our requirements. Consequently, a content driven audio search technology that is closely tied to our everyday experiences is emerging [2]. By examining songs, we can leverage information from these tracks for audio search. This approach captures the rhythms, melodies, frequencies, durations, and tones of the songs to aid our information retrieval [3]. This paper will concentrate on the principal melody of songs and explore it in depth [4]. Music consists of numerous distinct components, including songs, beats, chords, colors, dynamics, tempo, and rhythm.

Their interconnections primarily arise from rhythm, and these factors are essential to our auditory perception. However, there are currently no established standards to define these elements. MR technology categorizes music into three distinct types: monophonic, biphasic, and triphonic [5]. Biphasic music comprises two independent syllables [6], originating from different instruments (such as songs, guitar, bass), or two distinct beats (such as piano). Biphasic music is composed of two independent beats, each with a separate rhythmic element. By utilizing beat extraction technology, we can discern intricate pitches with varying tones and accurately forecast the major and minor tonalities of each sound. Melody serves as the central element of music, aiding individuals in grasping the emotions and concepts conveyed in the piece, and plays a vital role in music composition. The main melody, as defined by Poliner and colleagues, is a complex musical piece that evokes emotions, encapsulating the work's feelings and ideas. Melody is a key component of music, and its extraction is critical for effective audio retrieval. Thus, research into vocal melody extraction holds significant practical importance.

2. Early Work

Music melody extraction technology can be applied across a range of settings, from score recognition and ethnomusicology to pitch analysis, song feature and model examination, and the creation of electroacoustic music, all of which stand to benefit. Additionally, it can be utilized extensively in humming, cover song production, music genre classification, vocal suppression, and accompaniment generation, as well as singer recognition applications, becoming an essential preprocessing step in the music domain. Adobe's song extraction technology has been incorporated into Adobe Audition's audio editing capabilities. However, due to the confidentiality of associated commercial entities, the specifics of this technology remain undisclosed. Furthermore, there is a novel application in the MIR domain currently undergoing research. In the realm of music, utilizing different segments of the same song may result in varying styles. Typically, such practices may lead to plagiarism or other inappropriate actions [7]. Therefore, monitoring different sections of the same song could undermine its copyright protection and potentially lead to piracy.

Leng's findings have notably enhanced the harmonic regulation of audio signals. He adeptly converted the sequence of notes into MIDI format by utilizing a harmonic salience index, thus more effectively capturing the melody of the music. This accomplishment not only heralds the advent of a novel technology for detecting music melodies but also significantly advances the field of music research. A scholar fitted an observed spectrum to a specific pitch through the application of a pitch model [8, 9]. A critical challenge arises when using the pitch salience function to estimate the melody's fundamental frequency (f0), as it may only yield integer or fractional multiples of the true *fo* [10]. Furthermore, this algorithm typically estimates the approximate fo rather than the precise f0, leading to octave discrepancies.

Dressler identified high noise waveforms, commonly referred to as false highs, through detection methods.

Reddy discovered that when the average amplitude of the waveform spectrum of *f* exceeds that of the waveform spectrum of k·fo, it indicates that it is not genuine, but rather a fictive melody [11]. Hence, he recommended employing more precise measurement techniques, which significantly enhanced the detection accuracy of octave discrepancies [12]. These techniques can effectively mitigate irregularities in the true melody note spectrum envelope, thereby considerably decreasing the incidence of octave errors [13]. Although the methods mentioned earlier may offer certain enhancements, substantial pitch shifts during the pitch tracking phase could lead to more pronounced effects. Additionally, when background noise is present in the music, octave errors may arise, and a practical solution is currently lacking.

3. Principles of Vocal Melody Extraction Algorithm

3.1 Signal Preprocessing

Singing is a complex and ever-evolving speech signal. To more accurately capture the information it conveys, a series of preprocessing measures must be undertaken, including downsampling, normalization, framing, windowing, and Short Time Fourier Transform (STFT), among others. To reduce computational load in subsequent processing, all audio signals are down sampled to 8 kHz to better reflect their variations. By normalizing the input signal, we can maintain consistency in further processing. The specific formula is presented as follows:

$$\omega(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)] & 0\le n\le N\\ 0 & else\end{cases} \quad (1)$$

In this formula, *N* represents the window length, the frame length is 40 milliseconds, and the frame shift time is 20 milliseconds. If necessary, the Short Time Fourier Transform (STFT) can be used for time frequency analysis.

3.2 Note Segmentation and Voiced Segment Detection

The essence of music lies in its rhythm and melody. By using the DIS algorithm, we can divide the rhythm and melody in music into different parts, and use these parts to estimate their pitches. This method can provide more detailed and reliable pitch information. By using a short oscillation range, we can segment the syllables. We can measure the DIS metric difference between two data windows by sliding the data window of the t^{th} frame.

$$DIS(t)=\frac{\left(\boldsymbol{\mu}_{t,1}-\boldsymbol{\mu}_{t,2}\right)^{\mathrm{T}}\left(\boldsymbol{\mu}_{t,1}-\boldsymbol{\mu}_{t,2}\right)}{tr\left(\sum_{t,1}\right)+tr\left(\sum_{t,2}\right)} \quad (2)$$

In this formula, *u*. and *u* ..*z* represent the mean of two sections of audio features, while *tr* (2,1)and *tr* (,*z*) represent the trajectories of two sections of audio features, respectively.

After obtaining the DIS metric distance curve, we can determine the note segmentation point by observing the highest value of DIS(t). The specific filtering method is to compare the highest dis value with other values. If the highest value of dis is larger than the different values, it is used as the note segmentation point; otherwise, it is discarded.

In the music signal, we found some obvious parts, but they do not contain the fundamental frequency.Therefore, we first need to exclude these parts, then use the spectral variance method to determine the clear and unclear parts of the voiced portion.

3.3 Multiple Fundamental Frequency Estimation

We can effectively extract multiple fundamental frequencies from complex musical works, thereby improving their quality. If the start frequency of a noise segment is f, then the frequency of this noise segment is f.

$$Y(f) = \sum_{k=1}^{K} a_k \delta(f - kf_0) \qquad (2)$$

In this formula, *ak* is the coefficient of the kth harmonic. Based on different pitches, we can extract spectra corresponding to various peaks in $S(q)$ and form a set of benchmarks for use. In addition, since the pitch spectral range is generally 70~1000 Hz, we need to use equation (3) to estimate $S(q)$ accurately. By carefully selecting some frequencies with a higher oscillation range and utilizing their unique characteristics, we can effectively reduce the difficulty of calculation and improve the accuracy of testing.

Through the improved pitch salience calculation method, we can determine the salience value in the frequency range of 70~1000 Hz by detecting whether the amplitude of the log amplitude spectrum of the signal exceeds the threshold *T*. If it exceeds the threshold, we can use formula (3) to calculate the salience value, otherwise, we can set the salience value within this frequency range to 0.

By performing statistical analysis on the 70~1000Hz frequency band, the calculation complexity can be significantly reduced. Eventually, we can filter out the optimal three frequencies from numerous frequency bands that do not generate a harmonic or subharmonic relationship, thereby determining the optimal fundamental frequency. F represents the frequency distribution of the n^{th} candidate fundamental frequency, while s indicates the salience level of these fundamental frequencies.

3.4 Singing Melody Discrimination

The primary melody of vocal music is frequently referred to as the melody of human songs. When utilizing this method to extract notes, it may become apparent that they could originate from either songs or accompanying instruments. To ascertain whether these notes are part of the main melody of the song, it is necessary to implement a fundamental frequency discrimination model.

The specific network structure is shown in Figure 1.

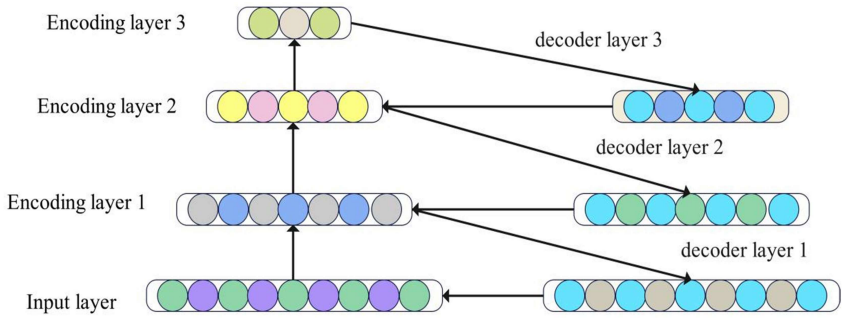

Figure 1. SAE Structure Diagram

By applying deep neural network technology, we propose a novel sound recognition approach that effectively pinpoints the fundamental frequency of sounds and is also applicable to speech signals. For this purpose, we will employ a sparse auto encoding neural network (SAENN) and a softmax classifier to develop a more precise sound recognition model.

In addition to the weight penalty parameter being defined as 0.0001, the desired sparsity is defined as p, the weight b is defined as 3, and the maximum number of iterations for the softmax classifier is limited to 500, while the total maximum number of iterations acceptable for the SAENN neural network is limited to 2000.

By using the L-BFGS algorithm, we can optimize the weights of the Hessian matrix during the training process. This method only uses data from m iterations, so it can build an approximation matrix of the Hessian matrix without adding additional burden. Furthermore, this method is highly robust.

4. Experimental Design and Result Analysis

4.1 Fundamental Frequency Discrimination Model Training and Testing

The audio sources used in this experiment are from Hsu's MIR-1K15, which includes 110 Chinese karaoke songs, of which 1000 have been carefully arranged with a sampling rate of 16 kHz, and also recorded 10 *ms* of vocal fundamental frequency, as well as two different channels, to present these wonderful works more fully.

Samples from 500 pieces of music were collected, of which some were used to train the SAE genome and the rest to evaluate the model's accuracy. In these checks, we will use consistent experimental conditions to assess the SAE's accuracy.

Using the L-BFGS method, we can train a neural network with sparse self numbering to detect a specific audio feature, thereby achieving high precision audio identification. Before this, we need to extract MFCC features from each sample and use them as independent labels, to obtain a reliable audio identification system that can effectively verify the correctness of these features. Through this experiment, we found that the recognition accuracy of this new model exceeded 85.1%, with a significant improvement in performance indicators, even surpassing traditional fundamental frequency judgement models. In addition, we found that using this new model reduces error rates in melody positioning and dramatically improves overall accuracy.

4.2 Music Main Melody Extraction Experiment

The two primary tasks involved in isolating the main melody from vocal music are: firstly, identifying the authenticity of the melody, including the error rate in melody positioning, the precision of the original pitch, the accuracy of the original chroma, and the overall correctness. The "Average Extraction Time (AET)" serves as a novel evaluation metric, aiding in comparing the computational complexity of enhanced algorithms with their original counterparts. AET quantifies the relationship between program runtime and the count of fundamental frequencies of the main melody, thereby providing insight into the algorithm's performance.

Utilizing Python as a base, we can develop a BP neural network comprising up to 200 neurons, with the potential to expand this to 400. Furthermore, by incorporating e-neurons, we can increase the total neuron count in both the input and output layers, thereby facilitating a more effective neural network architecture. After conducting 10 rounds of repeated trials, we ultimately achieved 100 valid instances. The variations in these validities are illustrated

in Figure 2.

As depicted in Figure 2, when the hidden layer of the BP neural network contains 192 neurons, it is evident that the hidden layer of the sparse deep neural network has a lower total neuron count. Nevertheless, through 10 trials, its accuracy markedly surpasses that of other BP model types. When the hidden layer neuron count of the BP neural network reaches 400, it outperforms other model types in terms of accuracy, achieving even better performance across 10 trials. However, while a limited number of hidden layers may enhance model accuracy, model distortion frequently persists.

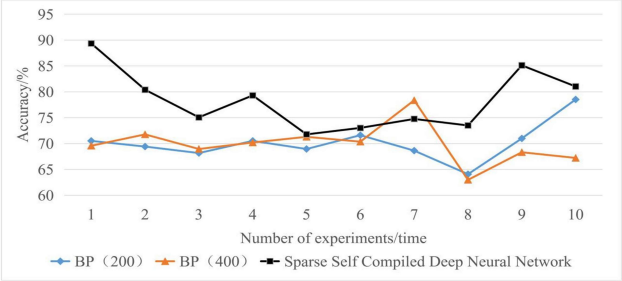

Figure 2. Accuracy of 10 Repeated Experiments with Various Algorithms

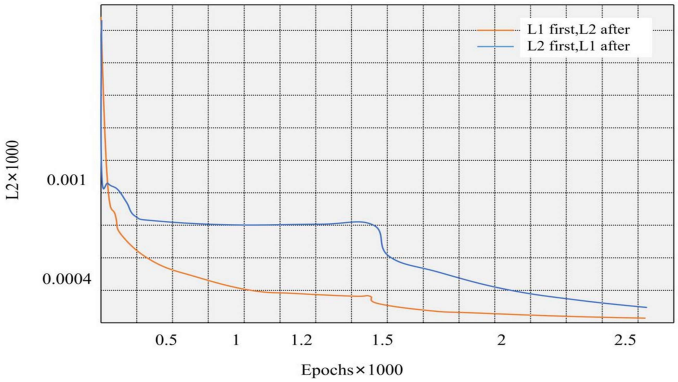

Figure 3. Relationship Between Loss Function and Number of Iterations

According to Figure 3, the new algorithm significantly outperforms the old one. On the contrary, the false alarm rate (VFAR) for melody localization has slightly decreased by 2%, indicating that the new algorithm can more accurately recognize the melody of accompaniments and songs. This is mainly due to the precision of the latest fundamental frequency analysis model. The optimized AET technology significantly shortens the processing cycle by 0.12 seconds, indicating that this technology can effectively reduce the difficulty of processing the pitch saliency function, thus better capturing the highlights in the music.

5. Conclusion

We introduce a novel technology that leverages sparse natural language processing methods to handle pitch saliency functions, alongside deep learning techniques for processing these functions. This approach enables faster processing of pitch saliency functions and improved pitch management. Additionally, we employ a cutting edge technique for processing pitch saliency functions, enabling speedier handling and better management of pitch. Ultimately, through model optimization, we can significantly decrease the false alarm rate in melody localization, thereby enhancing the model's overall accuracy. By refining the model, we can

reduce the average extraction time to 0.12 seconds and elevate the model's overall accuracy to 1.51%. This is a significant improvement than the existing conditions.

References

[1] Volberda, H. W., Khanagha, S., Fuller, Baden., C., et al. (2021). Strategizing in a digital world: Overcoming cognitive barriers, reconfiguring routines and introducing new organizational forms. *Long Range Planning,* 54 (5), 102110.

[2] Alduán, M., Sánchez, F., Álvarez, F., et al. (2011). System architecture for enriched semantic personalized media search and retrieval in the future media internet. *IEEE Communications Magazine,* 49 (3), 144-151.

[3] Nabati, M., Behrad, A. (2020). Multi-sentence video captioning using content oriented beam searching and multi stage refining algorithm. *Information Processing Management,* 57 (6), 102302.

[4] Zheng, W. E. N., Xin, Q. I., Keping, Y. U., et al. (2019). Content oriented common IoT platform for emergency management scenarios, 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC). *IEEE,* 1-6.

[5] Kojima, S., Kao, M. H., Doupe, A. J., et al. (2018). The avian basal ganglia are a source of rapid behavioral variation that enables vocal motor exploration. *Journal of Neuroscience,* 38 (45), 9635-9647.

[6] Gao, Y., Zhu, B., Li, W, et al. (2019). Vocal melody extraction via dnn based pitch estimation and salience based pitch refinement, CASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). *IEEE,* 1000-1004.

[7] Salamon, J., Gómez, E., Ellis, D. P. W, et al. (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine,* 31 (2), 118-134.

[8] Salamon, J., Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE transactions on audio, speech, and language processing,* 20 (6), 1759-1770.

[9] Gao, Y., Zhang, X., Li, W. (2021). Vocal Melody Extraction via HRNet Based Singing Voice Separation and Encoder-Decoder-Based Fo Estimation. *Electronics,* 10 (3), 298.

[10] Soltani, Z., Sørensen, K. K., Leth, J., et al. (2022). Fault detection and diagnosis in refrigeration systems using machine learning algorithms. *International Journal of Refrigeration* 144, 34-45.

[11] Hoq, M., Uddin, M. N., Park, S. B. (2021). Vocal feature extraction based artificial intelligent model for Parkinson's disease detection. *Diagnostics,* 11(6), 1076.

[12] Hui, X., Deying, L., Yan, L. (2009). Fault diagnosis for variable air volume systems using fuzzy neural networks, 2009 4th International Conference on Computer Science Education. *IEEE, 183-188.*