Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 3, 2025

Voiced and Unvoiced Sound Detection Based on Wavelet Transform

Mengyu Liu

Southwest Minzu University, Sichuan 610225, China

Abstract: Voiced-unvoiced detection is a fundamental research topic in speech signal processing, holding significant importance for applications such as speech recognition. Addressing the differences in frequency-domain energy distribution between voiced and unvoiced sounds, this paper proposes a voiced-unvoiced detection method based on discrete wavelet transform. By performing multi-level wavelet decomposition on speech signals, the method extracts energy features from high-frequency wavelet coefficients at different scales. Combined with an energy threshold decision strategy, it achieves the distinction between voiceless and voiced consonants. Experimental results demonstrate the method's effective differentiation between voiceless and voiced consonants, showing feasibility under the selected experimental conditions.

Keywords: Speech Signal Processing, Wavelet Transform, Voiceless-voiced Consonant Detection, Energy Features, Matlab.

1. Introduction

Voiced-unvoiced detection is a fundamental research topic in speech signal processing, playing a crucial role in applications such as speech recognition, speech enhancement, and speech coding. Speech signals are typically categorized into voiceless and voiced sounds: voiced sounds exhibit distinct quasi-periodic characteristics in their time-domain waveforms, with energy predominantly concentrated in lower frequency ranges; voiceless sounds, conversely, display non-periodic characteristics and contain relatively more high-frequency components. Due to the significant differences in time-domain and frequency-domain properties between voiced and voiceless sounds, their effective differentiation enhances the overall performance of subsequent speech processing algorithms.

Traditional voiceless/voiced detection methods primarily rely on features such as short-term energy, zero-crossing rate, and short-time Fourier transform (STFT). The short-term energy method distinguishes voiceless and voiced sounds by analyzing energy variations within brief intervals. While simple to implement and computationally efficient, it exhibits limited robustness in noisy environments and is susceptible to background noise interference. The zero-crossing rate method exploits differences in zero-crossing characteristics between voiceless and voiced sounds. While sensitive to high-frequency components, it is prone to misclassification in cases of weak voiced sounds or complex environmental conditions. Methods based on the short-time Fourier transform characterize the spectral properties of speech signals from a frequency-domain perspective. However, constrained by fixed window functions, they involve a trade-off between temporal and frequency resolution, resulting in insufficient adaptability to non-stationary speech signals.

To more effectively analyze non-stationary speech signals, wavelet transform has been introduced into speech signal processing due to its multi-resolution time-frequency analysis capabilities. Wavelet transform enables local analysis of signals at different scales. By performing discrete wavelet decomposition on speech signals, wavelet coefficients across various frequency bands can be obtained, reflecting the energy distribution characteristics of speech signals at each scale. In recent years, some studies have employed features such as wavelet packet decomposition, wavelet energy, or energy entropy, combined with threshold decision or simple classification strategies, to achieve voiceless/voiced consonant detection. This has enhanced the analysis capability for non-stationary speech signals to a certain extent.

Based on the above analysis, this paper proposes a voiceless-voiced consonant detection method based on discrete wavelet transform. This method effectively distinguishes voiceless from voiced consonants by extracting energy features from high-frequency wavelet coefficients at multiple scales and combining them with a hierarchical energy threshold decision strategy. Experimental results demonstrate that this method can effectively perform voiceless-voiced detection. It features simple implementation and low computational complexity, showing potential for application in speech signal processing. In terms of methodology, this approach belongs to the category of voiceless-voiced detection based on time-frequency analysis feature extraction combined with threshold decision strategies, emphasizing simplicity and feasibility in engineering implementation.

2. Wavelet-Based Voiced and Unvoiced Sound Detection Method

(1) Speech Signals and Voiced/Unvoiced Characteristics

Speech signals are typical non-stationary signals whose statistical properties change over time. Based on the vibrational state of the vocal cords during articulation, speech signals are generally categorized into voiceless and voiced sounds. Voiced sounds originate from periodic vocal cord vibrations, exhibiting distinct quasi-periodic characteristics in their time-domain waveforms with energy predominantly concentrated in lower frequency ranges. Voiceless sounds, typically formed by airflow through the vocal tract without significant vocal cord vibration, display non-periodic waveforms featuring relatively dispersed spectral distributions and richer high-frequency components.

Due to the differing mechanisms of voiceless and voiced sounds, they exhibit distinct characteristics in both the time and frequency domains. In speech signal processing,

voiceless and voiced consonants are often distinguished by extracting features such as energy and spectrum. Voiced consonants typically exhibit higher short-term energy and stronger low-frequency components, while voiceless consonants have relatively lower energy with a larger proportion of high-frequency components. Therefore, analyzing the energy distribution characteristics of speech signals from a frequency scale perspective provides an effective basis for voice and unvoiced consonant detection.

(2) Discrete Wavelet Transform and Energy Feature Extraction

The wavelet transform is a multiresolution time-frequency analysis method capable of performing local analysis of signals at different scales, making it particularly suitable for processing non-stationary signals. Compared to the traditional short-time Fourier transform, the wavelet transform can adaptively adjust its temporal and frequency resolutions according to the analysis scale, offering advantages in analyzing the transient characteristics of speech signals.

This paper employs discrete wavelet transform for multi-level decomposition of speech signals. Through a set of low-pass and high-pass filters, the original speech signal is successively decomposed into approximation coefficients and detail coefficients at different scales. The approximation coefficients primarily reflect low-frequency signal information, while the detail coefficients contain high-frequency components. Through multi-level decomposition, representations of the speech signal across different frequency bands can be obtained, enabling analysis of multi-scale features.

In voiceless-voiced consonant detection, the energy distribution of high-frequency components plays a crucial role in distinguishing between voiceless and voiced consonants. This paper selects the db6 wavelet as the basis function and performs a three-level discrete wavelet decomposition on the speech signal. At each decomposition level, the energy of the corresponding high-frequency wavelet coefficients is computed as a feature parameter. By incorporating multi-scale high-frequency energy features, the energy variations across different frequency ranges of the speech signal are more comprehensively captured, providing a reliable basis for subsequent voiceless-voiced judgment.

(3) Voiceless-Voiced Judgment Method

Building upon the feature extraction, this section presents the corresponding voiceless-voiced judgment strategy. Based on the aforementioned wavelet energy features, this paper employs an energy threshold-based decision strategy for voiceless/voiced detection. This method fully leverages the differences in high-frequency energy distribution across scales between voiceless and voiced consonants. By integrating multi-layer feature information for comprehensive decision-making, it enhances the stability of detection results.

The specific workflow is as follows: Preprocess the input speech signal and perform multi-layer decomposition using the discrete wavelet transform to obtain high-frequency wavelet coefficients for each layer; Then, the short-time average energy formula is applied to calculate the energy values of high-frequency coefficients at each decomposition layer, describing the energy distribution characteristics of the speech signal at different scales; Finally, based on experimental experience, corresponding decision thresholds are set for the high-frequency energy of each layer. The decision thresholds adopt a proportional threshold form based on energy features, with their specific value ranges and selection principles further explained in the experimental section.

The calculation formula for short-time average energy is as follows:

$$E_i = \frac{1}{N} \sum_{n=(i-1)N+1}^{iN} x^2(n)$$

Here, denotes the average energy within the i-th time interval, represents the value of the speech signal at the n-th sample point, and indicates the number of sample points contained in each time interval.

During the decision process, the current speech segment is classified as voiced or unvoiced by comparing the relationship between the high-frequency energy at different layers and the corresponding threshold. When the high-frequency energy is low, it is classified as unvoiced; while higher high-frequency energy indicates a voiced segment. By integrating multi-scale high-frequency energy information, this approach reduces misclassification risks associated with single-scale features and enhances the stability of voiceless/voiced detection. The method features a straightforward implementation with low computational complexity, making it suitable for both offline analysis and real-time processing of speech signals. The combination of multi-scale wavelet energy features with threshold-based decision-making enables effective differentiation between voiceless and voiced segments.

3. Experimental Results and Analysis

(1) Experimental Environment and Methods

To validate the effectiveness of the proposed voiceless-voiced consonant detection method based on discrete wavelet transform, experimental analysis of speech signals was conducted on the MATLAB platform. Single-channel speech signals were selected as the research subject, capable of simultaneously containing distinct voiceless and voiced segments to ensure the representativeness of the experimental results.

During the experiment, the speech signal was first read and preprocessed, followed by multi-level decomposition using the discrete wavelet transform. The db6 wavelet was selected as the basis function, and the speech signal underwent three-level discrete wavelet decomposition to obtain high-frequency wavelet coefficients at different scales. Subsequently, the energy values of the high-frequency wavelet coefficients at each decomposition level were calculated to characterize the energy distribution properties of the speech signal across different frequency scales. In the voicing/non-voicing determination phase, corresponding decision thresholds were set for the high-frequency energy of each layer based on the method proposed earlier. These thresholds were configured by introducing a proportional factor to the energy features, with the proportional factor ranging from 0.1 to 0.5. Through experimentation, a proportional factor of 0.3 was selected to balance detection stability and decision sensitivity.

(2) Voiced/Unvoiced Judgment Results

During the decision process, the speech signal is classified as voiced or unvoiced by comparing the relationship between high-frequency energy at different scales and the corresponding thresholds. When high-frequency energy is

low, it is classified as unvoiced; when high-frequency energy is high, it is classified as voiced. This decision strategy comprehensively utilizes high-frequency energy information across multiple scales, avoiding misclassification issues caused by single-scale features.

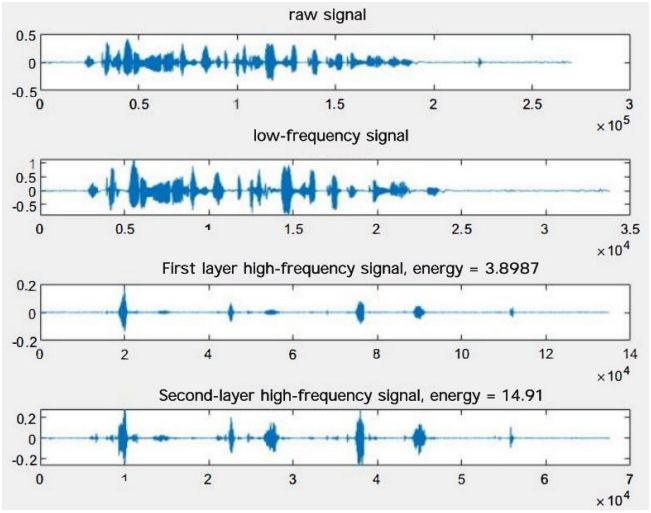

Figure 1. High-frequency and low-frequency signals and wavelet coefficient energy obtained after decomposing the original signal

Figure 1 shows the original speech signal along with its low-frequency and high-frequency components obtained through wavelet decomposition, along with the corresponding wavelet coefficient energies. The figure reveals distinct differences in energy distribution across different phonetic segments of the speech signal. Voiced segments typically exhibit higher energy levels, while the energy in voiceless segments shows relatively smoother variations.

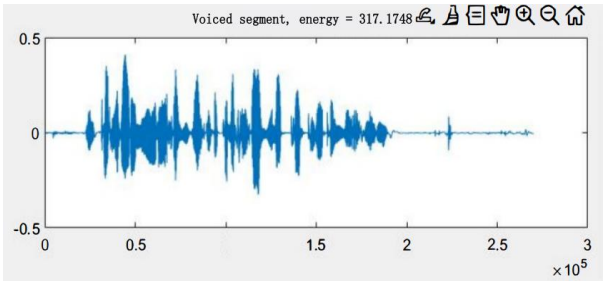

Figure 2. Identified as a voiced segment and calculated energy

Figure 2 displays the speech segments identified as voiced sounds in the voicing judgment results and their corresponding energy distributions. It can be observed that the segments classified as voiced exhibit higher high-frequency energy characteristics across multiple scales, indicating that the proposed method effectively reflects the differences in energy distribution between voiceless and voiced sounds.

(3) Results Analysis and Discussion

Experimental results demonstrate that the proposed clear-voiced/voiced detection method based on discrete wavelet transform effectively distinguishes between clear-voiced and voiced sounds. By incorporating multi-scale high-frequency energy features combined with a threshold-based decision strategy, the detection results generally align with the actual phonetic characteristics of speech signals.

Furthermore, this method features a straightforward implementation, low computational complexity, and minimal preprocessing requirements for speech signals, making it suitable for both offline analysis and real-time processing scenarios. It should be noted that the selection of threshold parameters influences detection outcomes; further optimization can be achieved by adjusting these parameters for different speech contents or noisy environments.

(4) Parameter Sensitivity Analysis

In energy-threshold-based voiceless/voiced consonant detection methods, threshold parameter selection influences judgment outcomes. To analyze the impact of threshold parameter variations on detection performance, this study compares detection results under different threshold ratio factor values.

In the experiment, the proportional factors of the high-frequency wavelet energy threshold were set to 0.2, 0.3, and 0.4, respectively, and the same speech signal was subjected to voiceless/voiced detection. When the scaling factor is small (e.g., 0.2), the decision threshold is relatively low, making the system sensitive to high-frequency energy changes. Some segments with lower energy may be misclassified as voiced sounds, leading to false positives. When the scaling factor is large (e.g., 0.4), the decision threshold increases accordingly, potentially misclassifying some low-energy voiced segments as unvoiced, leading to missed detections.

In contrast, when the scaling factor is set to 0.3, the decision results for voiced and unvoiced segments remain generally stable, effectively reflecting the energy variation characteristics across different articulation phases in the speech signal. Experimental results indicate that adjusting the threshold ratio factor within a reasonable range achieves a favorable balance between detection sensitivity and classification stability. The above analysis demonstrates that the proposed method exhibits stability within a reasonable threshold range. Thresholds can be appropriately adjusted based on specific speech signal characteristics and application scenarios to achieve more optimal voiceless/voiced consonant detection results.

4. Conclusion

This paper investigates and implements a voicing detection method based on discrete wavelet transform (DWT) for speech signals. Leveraging the differences in frequency characteristics and energy distribution between voiceless and voiced consonants, the method employs multiscale decomposition of speech signals via the discrete wavelet transform. It extracts energy features from high-frequency wavelet coefficients at different scales and combines these with an energy threshold decision strategy to achieve effective discrimination between voiceless and voiced consonants.

Experimental analysis of speech signals conducted on the MATLAB platform demonstrates that this method can effectively distinguish between voiceless and voiced segments in speech signals. The judgment results align with the overall trend of actual speech articulation characteristics. Experimental analysis and parameter sensitivity results indicate that, with appropriately set threshold parameters, this method achieves a good balance between detection sensitivity and judgment stability.

Compared to traditional voicing detection methods based on short-term energy or zero-crossing rate, this approach introduces a multiscale analysis mechanism in feature extraction. It fully leverages the multiresolution analysis advantages of wavelet transform, enhancing the characterization capability of non-stationary speech signals while maintaining a simple algorithm structure and low computational complexity, demonstrating practical value. It

should be noted that further improvements are possible in threshold selection and adaptability to complex noise environments. Future research may explore adaptive threshold strategies or integrating additional speech features to optimize the voiceless/voiced consonant detection method, thereby enhancing its adaptability in challenging application scenarios.

References

[1] Yan Fang. Design of a Wavelet-Based Speech Enhancement Algorithm [D]. Xiangtan University, 2021.

[2] Liu Yujun, Xia Cong. Application of Wavelet Analysis in Speech Recognition [J]. Cybersecurity Technology and Applications, 2014, (08):61-62.

[3] Zhu Fangyuan. Ma Zhiqiang, Chen Yan, et al. Research Review on Speaker Adaptation Methods in Speech Recognition [J]. Computer Science and Exploration, 2021, 15(12): 2241-2255.

[4] Wang Lianzi, Li Zhongxiao, Chen Qianqian, et al. Research on Voiced/Voiceless Judgment of Speech Signals Based on K-SVD Algorithm and Composite Dictionary [J]. Journal of Qingdao University (Engineering Technology Edition), 2020, 35(02): 17-23.

[5] Chen Xiaoli. Research on Fundamental Frequency Detection Algorithms for Noisy Speech [D]. PLA Information Engineering University, 2007.

[6] Zhu Jianrong, Zhu Jianping, Luo Ganghao, et al. Classification Methods and Performance Evaluation of Speech Denoising Techniques [J]. Modern Information Technology, 2025, 9(16): 1-7+14.