Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230891

Design and Implementation of Speech Signal Feature Extraction Algorithm Based on MATLAB

Jianli YANG[1], Long YANG, Yin CHEN
Chongqing University of Technology, Chong Qing, CHINA 400056;

Abstract. Speech signal is a widely-used type of information, with the development of artificial intelligence, the processing algorithms of speech signals is also rapidly evolving. The key to speech signal processing is to extract the feature information of speech signals. This paper extracts the fundamental period and LPC parameters of speech, and detects the resonance peak information of speech signals by focusing on real-time collected speech signals and related preprocessing. Based on understanding the principles of typical algorithms, the programming implementation and algorithmic results are provided, which can be applied to many places where speech signals are used, such as speech compression, speech recognition, and speech synthesis.

Keywords: Speech signal; Feature extraction; Fundamental period; Linear prediction; Resonance peak estimation

1. Introduction

Speech is a primary means of communication and serves as the basis for interaction between humans and computers as well as among computers themselves. The design and implementation of speech signal feature extraction algorithms can be applied to fields such as artificial intelligence, machine learning, and intelligent speech recognition. Analysis of speech information includes acoustic modeling, feature analysis, and compression coding. These technologies play a crucial role in areas such as speech compression coding, speech synthesis, speaker recognition, etc. [1]. For instance, speaker recognition, emotion recognition, speech command recognition, audio retrieval, audio compression, and decompression are all important applications of speech signal feature extraction. Moreover, these applications can also be utilized in fields such as criminal investigation, security access, target tracking, anti-counterfeiting, etc., providing people with better services and convenience while safeguarding their life and property.

2. Fundamentals of Speech Signal

2.1. Acoustic Model of the Signal

During speech production, the state of the vocal tract varies depending on the type of

[1] Corresponding Author: Jianli YANG, Chongqing University of Technology; e-mail: 18983350872@189.cn

speech, which can be classified into two categories: vowel and consonant sounds[2]. There are mainly two approaches for vocal tract modeling: one is to connect several pipes with certain cross-sections together to form a vocal tract system, while the other is to regard the channel as a resonance cavity and represent the resonance frequency of the cavity using resonance peaks. Due to the similarity of Cochlear organs in the human ear, the localization of its hair cells is related to their perception of frequency. Therefore, the resonant peak pattern is widely used and proven to be very useful. It has been shown that most speech can be described with only three harmonics, while some complex consonants and nasal sounds require around five resonant peaks. Typically, the features of the resonant peaks can be described by (1) as shown below.

$$V(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-1}} \tag{1}$$

The value of p, which represents the order of the entire pole filter, is usually within the range of 8 to 12, and each pole corresponds to a resonant peak. Previous studies have indicated that the characteristics of each channel remain constant between 10 to 30 milliseconds, which is crucial for rapid speech detection.

2.2. Feature Analysis of Speech Signals

Before digital processing of audio and video, filtering processes such as anti-aliasing and anti-interference are performed. Afterwards, sampling and quantization are carried out to convert analog signals into digital signals. Then, the short-time windowing and frame shifting methods are used to obtain a shorter time sequence for endpoint detection of speech signals. Endpoint detection methods such as dual-threshold method, correlation method, and spectral entropy are based on distinguishing speech and silence, which enables effective analysis and processing of speech signals.

2.3. Acquisition and Preprocessing of Speech Signals

A MATLAB program was developed to collect and read speech signals through a computer-based bioengineering system, as shown in the algorithm flowchart of speech acquisition and reading depicted in Figure 2. During the sampling of conversation sound, the polynomial least squares algorithm was utilized to eliminate the impact of trend items and direct current components on signal processing. A low-pass filter was applied to filter the acquired speech to address the problem of the poor separation of 50 Hz power frequency AC sound. Both pre-emphasis and post-emphasis methods were employed to address the issue of the concentration of power spectra at low frequencies increasing with frequency (low at low frequencies and decreasing at high frequencies), resulting in difficulty propagating at high frequencies.

2.4. Analysis of MATLAB Simulation Results

The he.wav file was read using the recordblocking and plot functions. Time is represented in seconds on the horizontal axis, while normalized values are represented

on the vertical axis. The simulation plot for speech acquisition and reading is depicted in Figure 1.

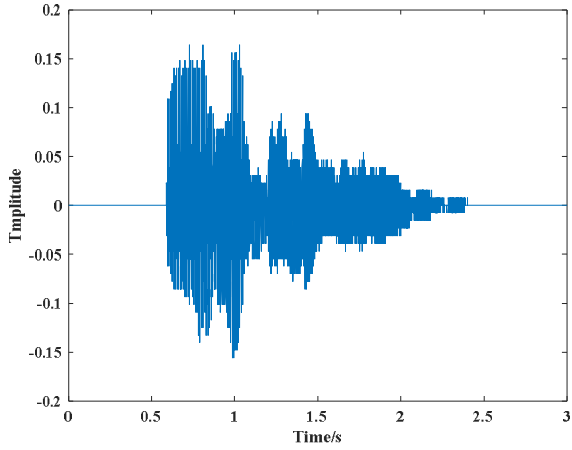

Figure 1. Simulation results for speech acquisition and reading.

The dual-threshold algorithm combines short-term energy and zero-crossing rate to detect silence, unvoiced sounds, and voiced sounds in speech signals. Short-term energy can distinguish between voiced sounds and silence but may misclassify unvoiced sounds, while zero-crossing rate can identify unvoiced sounds and silence in speech signals.

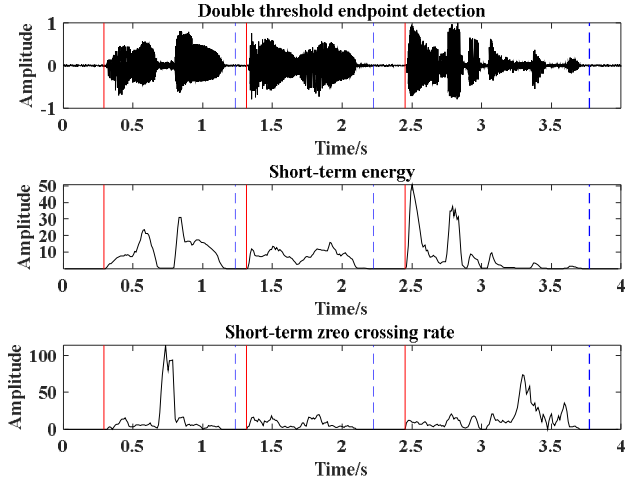

Figure 2. Simulation diagram of endpoint detection based on double threshold method.

The algorithm sets lower and upper thresholds based on two criteria. When the lower threshold is surpassed, it may be due to noise rather than the start of speech. Conversely, if the high threshold continues to exceed the low threshold, it indicates the start of a speech signal. Endpoint detection can be achieved through a two-step decision process.

The simulation diagram of endpoint detection based on the double threshold method is presented in Figure 2, which demonstrates the effectiveness of the algorithm.

3. Fundamental Frequency Extraction Algorithm and Implementation

3.1. Algorithms for Fundamental Frequency Estimation

Common methods for fundamental frequency estimation include peak detection, autocorrelation analysis, short-time Fourier transform, and linear prediction[3]. The peak detector analyzes the spectrum of the speech signal and estimates the fundamental frequency by analyzing the peak values of each spectral component. The autocorrelation analysis method utilizes the autocorrelation function within a specific region to estimate the fundamental frequency. The short-time Fourier transform method searches for the strongest spectral component within each time frame to determine the fundamental frequency of each frame. The linear prediction method provides higher prediction accuracy. These methods have broad applications in speech signal processing and recognition, and their results can be improved through the combination and optimization of multiple techniques.

3.2. Algorithm and Implementation based on Autocorrelation-based Fundamental Frequency Estimation

In order to digitally process speech signals, the speech signal is segmented into frames within a short period of time so that the signal in each frame can be analyzed using short-term autocorrelation. The expression for short-term autocorrelation function is shown in Equation (2).

$$\begin{aligned} R_{\mathrm{n}}(k) &= \sum_{m=-\infty}^{+\infty} X(\mathrm{m})\ W(n-m)X(m+k)W(n-m-k) \\ &= \sum_{\mathrm{m}=0}^{N-1-k} X(m+n)X(n+m+k) \end{aligned} \tag{2}$$

In this context, X（n）represents the speech waveform, W（n）denotes the window function, and N is the length of the window.

3.3. Analysis of MATLAB Simulation Results

Based on the short-time autocorrelation method, the fundamental frequency extraction can compare the similarity between two different time intervals, thus determining the fundamental frequency between the two intervals. In other different lag situations, the amplitude of the normalized lags is always less than one. The similarity between the two signals is highest when the lag is equal to one pitch period. Another method is to search for the maximum value between two maximum values and the difference between them is used to determine the initial value of a fundamental frequency cycle. The simulation graph of fundamental frequency extraction based on autocorrelation is shown in Figure 3.

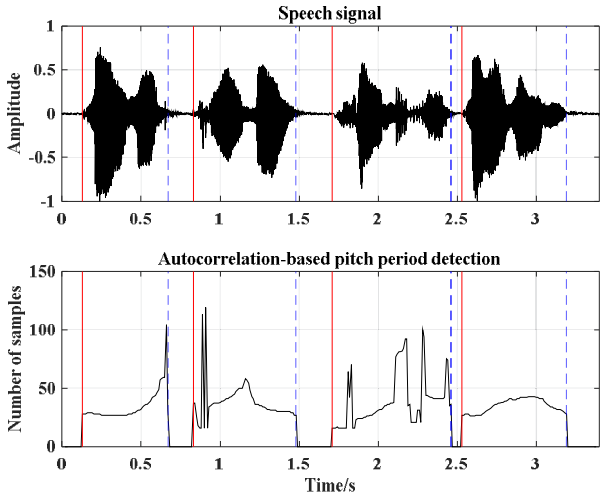

Figure 3. Simulation of gene detection based on autocorrelation.

4. LPC Parameter Extraction Algorithm and Implementation

4.1. Linear Prediction-Based Algorithms

Feature extraction of speech signals is a crucial step in speech signal processing. Among them, commonly used linear prediction-based algorithms mainly include traditional linear prediction method, Lyapunov predictive method, reverse spectral analysis, and weighted linear prediction-based method[4].

The traditional linear prediction method divides the speech data into a set of short-term speech data comprising multiple sampling points and models it to obtain two properties, namely, prediction error and prediction coefficient. Lyapunov predictive method is a dynamic, time-varying, and recursive method that can extract efficient features from unstable data, having a broader application domain. Reverse spectral analysis detects frequency periodicity through cepstral analysis and has good robustness for non-stationary signals, although it may generate significant biases in noisy environments. The weighted linear prediction-based method introduces weight factors to make the determination of prediction parameters more reasonable, enabling balancing of different components and improving the weighting coefficients of different components. The application of these algorithms can effectively extract the characteristics of speech signals and provide reliable support for subsequent tasks such as speech recognition and emotion recognition.

4.2. The algorithm implementation of linear prediction coefficient computation based on autocorrelation method.

The input parameter x represents the data of one frame, and p represents the linear prediction order. The output parameters are the predicted coefficients calculated according to the expression shown in Equation (3).

$$A\ (\mathrm{z}) = 1 + \sum_{i=1}^{p} a_i z^{-i} = \sum_{\mathrm{i}=0}^{p} a_i z^{-i} \tag{3}$$

The obtained predicted coefficients a_i consist of $p+1$ elements, with the first element always being 1. Meanwhile, e represents the minimum mean square error in the prediction computation.

4.3. Analysis of MATLAB Simulation Results

Based on the Levinson-Durbin autocorrelation method, a MATLAB function is developed and compared with the existing LPC function in MATLAB. The evaluation is conducted by comparing the calculation results of linear prediction coefficients based on the autocorrelation method. The simulation results of linear prediction coefficient computation based on autocorrelation method are presented in Figure 4 to demonstrate the testing performance.

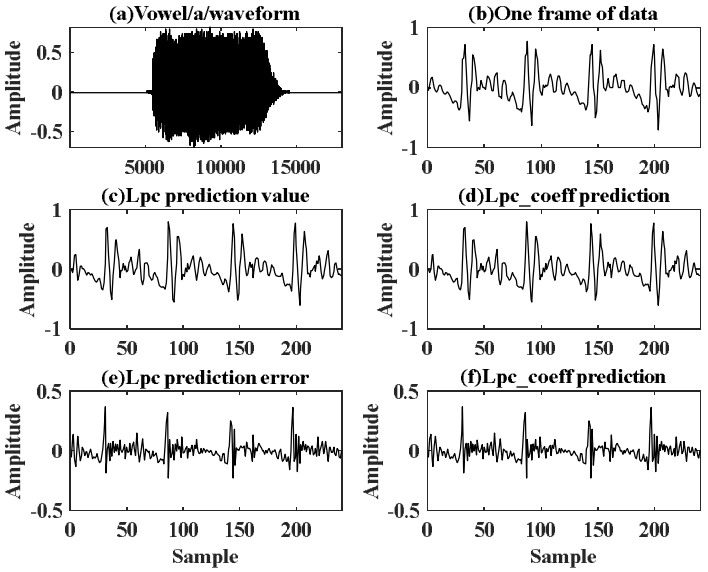

Figure 4. Simulation of gene detection based on autocorrelation.

5. Resonant Peak Detection Algorithm and Implementation

5.1. Resonant Peaks of Speech Signals

Resonant peak is a crucial feature parameter in speech signal processing, which represents the most direct source of pronunciation information and reflects important characteristics of the resonant properties of the vocal tract. The key to extracting resonant peak parameters lies in estimating the spectral envelope of the speech signal, with the

maximum value in the envelope generally regarded as the resonant peak. Similar to fundamental frequency detection, resonant peak estimation seems easy at first glance, but is actually plagued by many problems.

5.2. The Algorithm and Implementation of Resonant Peak Estimation Based on Cepstrum Method

The cepstral method extracts features by transforming the spectrum of a signal into the time-domain representation of cepstral coefficients[5][6]. The fundamental idea behind the cepstral method is to apply a discrete cosine transform (DCT) or discrete Fourier transform (DFT) to the logarithmic spectrum of the signal, yielding cepstral coefficients. These coefficients represent the envelope information of the signal and can be utilized for acoustic feature extraction, estimation of resonance peaks, and training acoustic models, among other applications.

To better avoid false peaks, merging of resonant peaks, and high-pitched speech in the detection of resonant peaks, the cepstrum method is adopted to estimate the resonant peaks. The specific steps and principles are as follows:

(1) The speech signal $\mathrm{x}(i)$ is pre-emphasized and subjected to windowing and framing. Subsequently, Fourier transform is performed to obtain the expression $X_{\mathrm{i}}(k)$, as shown in equation (4).

$$X_{\mathrm{i}}(k)=\sum_{n=0}^{N-1} x_i(n)e^{-j2^{\pi kn/N}} \tag{4}$$

(2) The cepstrum of $X_{\mathrm{i}}(k)$ is obtained to yield the expression $\mathrm{x}'_i(n)$, as shown in equation (5).

$$\mathrm{x}'_{\mathrm{i}}(n)=\frac{1}{N}\sum_{k=0}^{N-1}\lg|X'_{\mathrm{i}}(k)|e^{j2\pi kn/N} \tag{5}$$

(3) The windowing of the cepstral signal $\mathrm{x}'_i(n)$ with a window $h(n)$ results in the expression $h_{\mathrm{i}}(n)$, as shown in equation (6).

$$h_i(n)=x'_i(n)\times h(n) \tag{6}$$

(4) The envelope of $h_{\mathrm{i}}(n)$ is obtained to yield the expression $H_{\mathrm{i}}(k)$, as shown in equation (7).

$$H_{\mathrm{i}}(k)=\sum_{n=0}^{N-1} h_i(n)e^{-j2\pi kn/N} \tag{7}$$

(5) By identifying the maximum value on the spectral envelope of the speech signal, the corresponding parameters of the formants can be extracted.

5.3. Analysis of MATLAB Simulation Results

The input parameter u represents a frame of speech input data. cepstL denotes the width of the window function on the inverse frequency axis. The output parameters Val, Loc, and spect represent the amplitude, location, and envelope, respectively, of the resonance peaks. Figure 5 depicts the simulation result of resonance peak estimation based on the cepstral method.

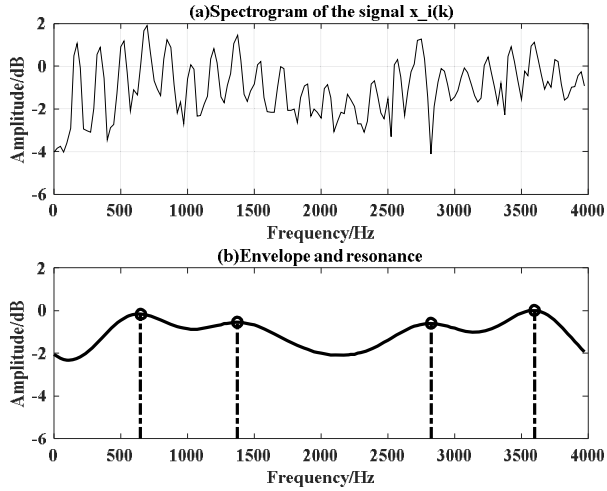

Figure 5. shows the simulation result of resonance peak estimation using the cepstral method.

6. Conclusion and Future Work

This paper introduced the extraction methods and principles of three feature parameters: pitch, linear prediction coefficients, and resonance peaks, which has fulfilled the requirements for using speech signals in practical scenarios and validated the functionality through testing and verification by implementing speech signal feature extraction algorithms in MATLAB. The study emphasizes the integration of research on speech signal extraction technology with practical applications, establishing a solid foundation for future development and industry collaboration. With the advancement of artificial intelligence technology, precision feature extraction and efficient algorithm implementation are the key directions for the continuous improvement and research of speech signal processing technology.

Research Project

1.Study on Intelligent Signal Perception Method for Optical Communication System based on Blockchain Technology (2022xzky06).

2.2022 National Project of Student Innovation and Entrepreneurship Training Prog ram(202212608006).

References

[1] Anna Mari Makela,Paavo Alku.Human cortical determined by speech fundamental frequency[C].Neurolmage.Cincinnait: IEEE,2002:1300-1305.
[2] Feng, Qi, et al. "End-to-end speaker recognition from raw waveforms with SincNet." 2019 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2019.
[3] Mirsamadi, Soheil, et al. "Automatic speech emotion recognition using recurrent neural networks with local attention." ICASSP 2017-2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.
[4] Z. Wang, M. Liu, X. Feng, and X. Wu, "A novel deep learning-based feature extraction method for speech signals," IEEE Access, vol. 8, pp. 40258-40267, 2020.
[5] Y. Wang, Y. Zhang, and L. Xu, "Robust deep feature extraction for speech recognition with adversarial training," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2626-2639, 2020.
[6] D. Liu, Y. Song, Y. Qin, and Z. Cao, "A novel speech feature extraction method based on deep belief networks and statistical covariance analysis," IEEE Transactions on Instrumentation and Measurement, vol. 68, pp. 1818-1827, 2019.