An Improved Speech Base Frequency Extraction Algorithm

Xiaojing Yu¹, Hongyan Zhang¹*, Yanmei Shao² and Yingying Jin¹

¹ Meigu College, Business School, Shandong Normal University, Jinan, China
² School of Chinese Language and Literature, Shandong Normal University, Jinan, China
Email: yuxjjw@163.com

Abstract. Traditional fundamental frequency extraction algorithms mainly include cepstrum method, short-time autocorrelation method. Some end-point extraction methods can’t correctly select the part of effective value due to the particularity of dialect audio and the noise influence. And, some fundamental frequency extraction methods cannot get the data consistent with the trend of tuning values. Against the above, this paper proposes an improved double threshold method for the extraction of the basic frequency according to the characteristics of the vowels of the initials of the one-word tone, which extracts the continuous vowels of the one-word tone, and looks for the basic frequency extraction algorithm that meets the classification of the tones of one-word. Finally, the effective basic frequency data is obtained. This method is more effective than other methods in speech tone recognition. It is of great significance to the study of speech tone, especially the study of dialect tone.

Keywords. Endpoint detection; double threshold method; fundamental frequency extraction.

1. Introduction

Chinese syllables are mainly composed of initials, finals and tones. Because of the composition of Chinese intonation’s different from that of other languages, it is an important content to study the characteristics tones of Chinese. The change of fundamental frequency determines the tone, which is a very important phonetic feature that can effectively distinguish semantics and emotions [1]. The fundamental frequency plays an important role in the study of speech features, which can be applied to all aspects of linguistics, such as tone recognition [2], emotion recognition [3], automatic music notation [4], speaker recognition [5].

Fundamental frequency extraction analysis methods can be divided into time domain, frequency domain, inverted frequency domain and other domain analysis methods according to the different extraction parameters. Signal processing methods in the time domain mainly include energy and average amplitude, average zero crossing rate, autocorrelation function [6]. The autocorrelation function method is easy to occurred frequency double error and half frequency error [7] under some conditions (such as low SNR). Aiming at this kind of problem, the basic frequency extraction algorithm of the autocorrelation function based on linear prediction [8] and the weighted autocorrelation algorithm based on wavelet packet transform [9] have appeared. Wavelet transform has the advantages of low computational complexity, suitable for transient data analysis, and easy implementation on hardware [10]. On this basis, there are proposed the algorithm of extracting the fundamental frequency based on wavelet transform [11] and the method of de-noising the fundamental frequency with wavelet transform [12]. Cepstrum method, which is to take logarithm of speech signal and then carry out inverse Fourier transform [7], has the good extraction effect of fundamental frequency in the case of no noise. There
have come up with a new cepstrum pretreatment method to enhance the extraction effect [13], noise removal, optimization method of improved algorithm [14], autocorrelation and cepstrum weighted square operation method [15], which have improved the robustness and accuracy of extraction.

The double threshold method is based on short time average energy and short time zero crossing rate. The principle is that the final part can be found by comparing the short time average energy, while the initial part can be found by using the short time average zero crossing rate. Existing improvement methods include the fusion of speech enhancement to improve the accuracy under the condition of low SNR [16], the method of adjusting the number of thresholds and smoothing filtering [17], replacing the short-term average zero crossing rate with a better spectral centroid feature [18].

In the aspect of the extraction of the basic frequency of the dialects, the traditional algorithm of the basic frequency extraction for mandarin cannot meet the requirements of the basic frequency data in the study of tone recognition, and cannot extract the basic frequency of the dialects as the same as the extraction of mandarin, which seriously hinders the study of Chinese dialects. The reason is that the tone value of dialects is more complex and the traditional extraction algorithm cannot distinguish the voiced part and the change of pitch frequency well.

This method aims to find a new algorithm to extract the basic frequency of a single word, which can meet the requirements of data. By using the data set of Linyi City to improve the algorithm experiment, a new method is proposed by improving the end point extraction in audio preprocessing and combining with the basic frequency extraction method of simplified inverse filtering, which can optimize fundamental frequency extraction. The basic frequency data extracted by this method is basically consistent with the known modulation value, and it can be judged that it can meet the requirements of the basic frequency data extraction.

### 2. Fundamental Frequency Extraction Process

The proposed method can be divided into two steps: endpoint detection and fundamental frequency extraction. The most used is the double threshold method in the traditional detection methods. Different from the traditional method, this method has been improved to reduce the influence of short-time zero crossing rate in the process of extraction, focus on extracting voiced segments and carrying out further operation.

![Figure 1. Steps of fundamental frequency extraction.](image)

The amplitude normalization and frame splitting of the data are carried out after the audio is converted into data. Through endpoint extraction, the vowel segments are found and intercepted so as to avoid invalid segments in the next operation. The intercepted data is divided into frames again and processed into the data needed for the next step (figure 1). After repeated attempts, the tuning trend of the basic frequency extracted by the simplified inverse filtering method is the best in the dialect corpus.

#### 2.1. Endpoint Extraction

Setting the length of a leading no-talk segment and giving the frame length and frame shift to obtain the number of silent frames. And calculate the energy value and zero crossing rate value of each frame respectively.

\[
amp = \sum_{m=1}^{N} x_i^2 (m) 
\]

\[
zcr = \sum_{m=1}^{N} |\text{sign}[\bar{x}_i (m)] - \text{sign}[\bar{x}_i(m - 1)]|
\]
Among equation (1) $x_i^2$ as piecewise function, when $|x_i^2|$ is bigger than $s$, its value is $x_i^2$, otherwise is zero, which $s$ for a very small number. In equation (2), $\text{Sign}(n)$ function is also piecewise, when $|n|$ greater than or equal to 0, the value is 1, otherwise to 1.

We obtain the average energy value $\text{AMP}$ and the average zero crossing rate value $\text{ZCR}$ of the leading no-talk segment. Set the thresholds $\text{amp}_1$, $\text{amp}_2$, $\text{zcr}_1$ according to these two values:

$$\text{amp}_1 = 3*\text{amp}$$  \hspace{1cm} (3)  \\
$$\text{amp}_2 = 6*\text{amp}$$  \hspace{1cm} (4)  \\
$$\text{zcr}_1 = 4*\text{zcr}$$  \hspace{1cm} (5)

The next to secondary judgment. The first-level judgment, which belongs to the rough judgment, compares the energy of each frame with $\text{amp}_2$, and the one greater than this value must be the sound segment. The second decision, voice frames that do not meet the first decision will be given a second decision. The energy and zero crossing rate of this frame are compared with $\text{amp}_1$ and $\text{zcr}_1$ (the value of $\text{zcr}_1$ is 0, which has no effect). If the conditions are met, it can be considered that it is in the speech segment. This decision can reduce the error caused by pronunciation problems. Statistical meet the conditions of continuous speech frame length, if more than one designated numerical 6, can be identified as speech segment, if it is not, the speech segment is considered to be null. Compare in turn, we will get the beginning and end of the paragraph.

It can be found that some of the detected segments are detected as a segment (figure 2), while some get multiple segments (figure 3), that requires us to effectively distinguish which segments are valuable. After some processing, it is cut out and further processing is carried out on the newly obtained fragments.

**Figure 2.** A picture results in a segment.  
**Figure 3.** A picture results in multiple segments.

### 2.2. Extraction of Basic Frequency by Simplified Inverse Filtering Method

This paper adopts the method of designing the screening algorithm, optimize and screen the obtained segments by observing the characteristics of audio segments, and obtain the effective segments for fundamental frequency extraction. Because of the different preprocessing methods required by fundamental frequency extraction and endpoint detection, so it is necessary to reprocess the audio. The data should be sampled down, and the frames and endpoints should be detected to ensure the accuracy of the audio segment. A simple inverse filtering algorithm is used to extract the fundamental frequency. Through the improvement, the endpoint detection and fundamental frequency extraction are combined to improve the detection effect of dialect speech.

**Figure 4.** Steps of simplified inverse filtering method [19].
The specific operation steps are as follows: The speech signal set as X is processed through a low-frequency bandpass filter to get X1. According to experience, a filter with a bandpass of 60~500 is selected. After that, X2 was obtained by sampling down according to 4:1. After frame splitting, LPC coefficient was extracted to calculate the autocorrelation function and then RU1 was obtained. RU was obtained by increasing the sampling rate by 4 times, and the pitch period of the speech was obtained by detecting the correlation peak value (figure 4).

3. Experiment
Three experiments were carried out for the method proposed in this paper.

3.1. The Data Set
The data set used in this experiment is the Linyi dialect data set, which is derived from the special data set of language protection, including the speech data of 8 counties in Linyi. In each region. There are more than 2000 words, vocalizations, sentences and other fragments recorded by several speakers. The recorder is Yanmei Shao, University of Shandong Normal University.

In this experiment, only the word fragments of elderly men were used for 8 different tonality values. The Linyi dialect database has been recognized by professionally trained dialect researchers, which meets the verification requirements of this experiment.

3.2. Endpoint Detection Experiment
As the national common language, mandarin is characterized by simple and tone clearly, which has the advantage of easy recognition in the aspect of voice feature extraction. Compared with dialects, mandarin is more suitable for all kinds of methods. As for the processing and extraction method of Linyi dialect pronunciation described in this paper, it should also be applicable to the processing of Mandarin. Therefore, this experiment is used to test the effect of this method on the extraction of mandarin. The mandarin database used in this experiment was recorded by professional recording personnel, and the standard mandarin was of reference significance.

Table 1. Accuracy of the method in this paper for mandarin.

|       | 1st tone | 2nd tone | 3rd tone | 4th tone |
|-------|----------|----------|----------|----------|
| Accuracy (%) | 97.9 | 84.7 | 81.5 | 96.6 |

The experimental results are shown in table 1. The proposed method has a good extraction effect on the first and fourth tones, with the highest accuracy of 97.9%, but for the second and third tones, the effect is slightly worse, with the lowest only reaching 81.5%.

3.3. Experiment with Different Tuning Values
The different characteristics of different tuning values, the adaptation effect of the method in this paper is also different. Therefore, for different tuning values, the accuracy of processing will not be exactly the same, so it is necessary to further obtain the accuracy of the results through experiments.

The audio frequency with different tuning values in some areas is selected and experimented with the method in this paper, and the accuracy rate in line with the trend of tuning values is obtained.

Table 2. Accuracy of different tuning values.

|       | Junan Accuracy (%) | Tancheng Accuracy (%) | Lanshan Accuracy (%) |
|-------|--------------------|-----------------------|----------------------|
| 55    | 70.7               | 44                    | 55                   | 67.0 |
| 53    | 66.7               | 51                    | 60.4                 | 69.4 |
| 213   | 66.7               | 213                   | 64.1                 | 65.6 |
| 312   | 75.6               | 24                    | 68.4                 | 67.4 |
| average | 69.9              | average               | 61.7                 | 67.4 |
The results are shown in Table 2. In general, the accuracy of tone can be more than 60%, among which the accuracy of Linshu County’s audio is the highest, reaching 70.1%. The accuracy of 53 tone value in Fei County is too low, which can be regarded as an error ignored. For a single tone value, the best value of 312 in Linshu is 77.8%, and most of them can maintain around 65%, which shows a good effect.

Experiments have shown that the accuracy of the same tone value is different in different regions (the accuracy of 44 tone of Linshu was 67.3% and that of Pingyi was 71.4%), and the accuracy of different tones varies from region to region (the accuracy of four tone values in Linshu is different). In general, the accuracy of the method described in this paper meets the requirements of basic frequency data, but it needs further improvement.

### 3.4. Compare Experiments with Different Methods

Traditional fundamental frequency extraction methods are mostly designed for Mandarin, but are not very suitable for dialect pronunciation. Compared with this, the method in this paper focuses on the basic frequency extraction of dialect, which is more targeted and has better effect.

In this experiment, the accuracy of the traditional method is compared with that of the proposed method, and each tone is compared using a variety of methods, such as the linear correlation cepstrum method, autocorrelation function method, cepstrum method and average amplitude difference method.

| Method in this paper | Linear predictive cepstrum method | Autocorrelation function method | Cepstrum method | Mean amplitude difference method |
|----------------------|----------------------------------|---------------------------------|-----------------|---------------------------------|
| 55                   | 70.7                             | 35.9                            | 18.3            | 36.5                            | 25.4 |
| 312                  | 75.6                             | 21.2                            | 20.0            | 30.2                            | 31.6 |
| 53                   | 66.7                             | 28.7                            | 18.6            | 34.3                            | 38.7 |
| 213                  | 66.7                             | 25.5                            | 17.2            | 28.9                            | 32.8 |
| 24                   | 68.4                             | 43.6                            | 41.8            | 46.2                            | 43.6 |
| 44                   | 71.4                             | 22.8                            | 19.5            | 29.3                            | 26.8 |
| 51                   | 60.4                             | 36.2                            | 17.4            | 23.6                            | 16.8 |
| 214                  | 65.6                             | 32.5                            | 32.6            | 28.1                            | 34.8 |
| 1st tone             | 97.9                             | 94.2                            | 87.2            | 88.9                            | 92.4 |
| 2nd tone             | 84.7                             | 77.3                            | 83.1            | 83.7                            | 84.9 |
| 3rd tone             | 81.5                             | 80.3                            | 84.2            | 82.0                            | 79.2 |
| 4th tone             | 96.6                             | 94.1                            | 86.0            | 91.4                            | 88.6 |

In Table 3, the results show that the proposed method is not significantly different from the traditional algorithm in the toning of Mandarin. The effect of the second and the third tone is similar, but the first tone is better than other methods. Overall, it is better than the traditional algorithm. Among the dialect tuning values, the results obtained in this paper are the best, which the accuracy rate is maintained above 60%. In the traditional method, the best result is only 46.2%, among which the autocorrelation function method is the worst. The other three algorithms have different effects for different tuning values. Due to the effect is too different and the accuracy is not high, they are worse than the method in this paper on the whole.
The tonal direction obtained by the traditional fundamental frequency extraction method has a larger error than that obtained by the label, that’s because the traditional endpoint extraction method and the fundamental frequency extraction algorithm are not suitable for the processing of dialect audio.

4. Conclusion
This paper proposes a fundamental frequency on the basis of the improved double threshold method is a new method, which combines the improved double threshold method and the simplified inverse filtering method to form a new method. Compared with other methods, this method to extract the fundamental frequency accuracy slightly high, but there is still room for improvement and needing further improvement. The experiment was carried out different accuracy experiments from three aspects, including comparison with mandarin, difference of different tone values and comparison of traditional methods. The results show that it is more effective than other methods in extracting the endpoints and fundamental frequencies of one-word dialects this method, and it is of great significance to the study of one-word tones, especially the study of dialect tones.

Acknowledgments
This work was supported by the Ministry of education of Humanities and Social Science project (Grant No. 19YJCZH244), the National Natural Science Foundation of China (Grant No. 61876101, 61802234, 61806114).

References
[1] Song Z Y 2013 Application of MATLAB in Speech Signal Analysis and Synthesis (Beihang University Press).
[2] Xu S L and Ying Y 1994 A multi-feature fuzzy recognition method for chinese tone Pattern Recognition and Artificial Intelligence (01) 60-5.
[3] Abdel-Hamid L 2020 Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features Speech Communication 122.
[4] Zhou R H and Yan Y H 2010 An approach to fast automatic polyphonic music transcription Acta Acustica 35 (2) 282-7.
[5] Zhang J P, et al. 2010 Lang span prosodic features for speaker recognition Acta Acustica 35 (02) 267-9.
[6] Hu H Y, Yuan M, and Feng H H 2012 A dynamic smoothing based real-time pitch detection algorithm Technical Acoustics 31 (06) 583-8.
[7] He J 2012 Research on Pit Detection Technology of Noised Speech Signal (University of Electronic Science and Technology of China).
[8] Bai J and Wei G 2005 A New Method for Speech Signals Pitch Detection Based on LPC and Autocorrelation Electroacoustic Technology (08) 43-46.
[9] Sun T T and Zhang X B 2017 A weighted autocorrelation method for pitch detection based on wavelet packet transform Computer Engineering & Science 39 (08) 1525-29.
[10] Zhang X, Liu B, and Li S G 2002 The implementation of wavelet transform in speech keynote period picking Communications Technology (02) 9-11.
[11] Dong L, Xin W R, and Ye D F 2005 Audio signal pitch extraction based on wavelet transform Electroacoustic Technology (06) 50-55.
[12] Fu T J 2015 Research on The Pick-up Algorithm of Phonetic Eigenvalue Based on The Wavelet Transform (Shenyang Ligong University).
[13] Hu, X., et al.,2013 A cepstrum-based preprocessing and postprocessing for speech enhancement in adverse environments Applied Acoustics 74(12).
[14] Zhang J K and Peng Y 2017 Improved algorithm of pitch detection based on cepstrum Electroacoustic Technology 41 (Z2) 113-116+130.
[15] Chen P D, Huang H and He L 2015 Improved algorithm for pitch detection based on ACF and CEP Computer Applications and Software 32 (01) 163-6.
[16] Liu Y Z and Tian J B 2016 Double threshold endpoint detection algorithm based on speech enhancement Measurement and Control Technology 35 (11) 33-35.

[17] Xue S Y 2015 Research on speech endpoint detection based on the improved dual-threshold algorithm Electronic Design Engineering 23 (04) 78-81.

[18] Jiang N 2020 Improvement of speech endpoint detection method in cases Journal of Criminal Investigation Police University of China (03) 123-8.

[19] Yang X J 1995 Speech Signal Digital Processing (Publishing House of Electronics Industry).