An Automatic Pitch Detection Method Based on Multi-feature for Mandarin Speech

Runshen Cai

Computer Science and Information Engineering College, Tianjin University of Science and Technology, Tianjin, China
crs@tust.edu.cn

Abstract

There are many traditional pitch detection methods, but most of them can't perform perfectly for different speakers, applications and environmental conditions. For this reason, a pitch detection method based on multi-feature is proposed. Firstly, the speech signals are pre-filtered. Secondly, the speech signal pre-filtered is segmented into syllables. Finally, the pitch period is obtained by wavelet transform and the maxima selected. Experiments show that this method can increase the performance of pitch detection in both clean and noisy environment compared with weighted wavelet method.

Keywords: *pitch detection; wavelet transform; speech signal; speech analysis; signal processing*

1. Introduction

Pitch detection is one of the most fundamental, important, and difficult task in the area of speech analysis, speech synthesis, speech coding and speech recognition. There are many traditional pitch detection algorithms such as the short-time autocorrelation function (ACF) [1], short-time average magnitude difference function (AMDF)[2], cepstrum (CEP) [3] and wavelet transform [4] algorithms. These methods are all based on short-time stationary speech, so they are not always useful for many different speech signals and many modified methods [5] are presented to solve the problems.

In recent years, the wavelet transform has been successful used in many speech processing applications [6]. Wavelet transform can analyze time-frequency characteristics of speech, and can track abrupt changes of speech. So it becomes a powerful tool for pitch detection. There are many modified pitch detection method based on wavelet transform have been developed and get some better performance in some cases.

But because there are a lot of problems in pitch detection, thus no one algorithm has been developed so far performing perfectly for all different speakers, applications and environmental conditions.

In order to accurately estimate pitch period in both clean and noise environment, this paper proposes a modified pitch detection method for noisy speech signals. The flow chart of this approach is shown in Figure 1.

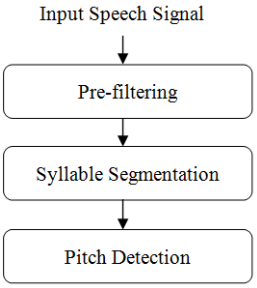

Figure 1. The Flow Chart of the Proposed Pitch Detection Mehtod

2. Pre-filtering

Abundant harmonic component and pitch frequency contained in speech signals usually is the main factor of affecting the performance of traditional pitch detection method. But for noisy speech, noise often has very bad effect and can't be neglected.

In order to decrease the influences from high frequency noise and high frequency formants on pitch detection, it is necessary to pre-filter noisy speech signals by a low-pass filter. According to the scope of pitch frequency of speech, a 5-order low-pass elliptic filter [7] whose cut-off frequency is 800Hz is used in this paper. The filter not only eliminates the influences from high frequency noise and main formants, but also reserves the first and the second harmonics when pitch frequency is less than 500Hz and pitch frequency of speech signal is always less than 500Hz. The transfer function of this filter is given by,

$$H(z)=\frac{(0.008233 - 0.004879z^{-1} + 0.007632z^{-2} + 0.007632z^{-3} - 0.004879z^{-4} + 0.008233z^{-5})}{(1 - 3.6868z^{-1} + 5.8926z^{-2} - 5.0085z^{-3} + 2.2518z^{-4} - 0.4271z^{-5})} \quad (1)$$

3. Syllable Segmentation

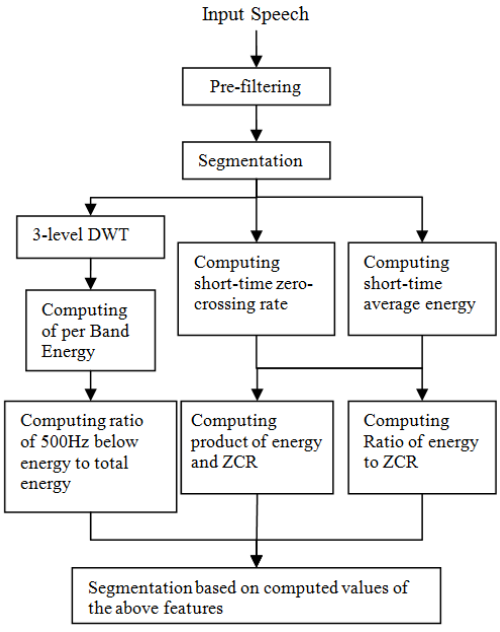

Figure 2. Flow Diagram of the Method

Figure 2, gives a flow diagram representation of the method.

First the input speech signal, sampled at 8 kHz, is denoised by pre-filtering. Then the signal is segmented into 30 ms long segments with 20ms overlap. After the segmentation stage, the following analysis and feature-extraction processes are implemented for each segment. Finally, syllable segmentation based on the computed features is performed.

3.1. Calculate Features

3.1.1. Short-time Average Energy: The average energy of the i-th speech signal segment, defined as (2):

$$E_i = \left(\sum_{n=0}^{N-1} \left| x_i(n)^2 \right|\right) \Big/ N \tag{2}$$

It provides a convenient representation that reflects the variations of the amplitude of the speech signal [8]. The average energy of non-speech segments is generally much lower than that of speech segments, and for speech segments, that of unvoiced segments is generally much lower than that of voiced segments. Furthermore, average energy is always becoming lower at the syllable boundary than in the syllable.

3.1.2. Short-time Zero-crossing Rate (ZCR): In the context of discrete-time signals, a zero-crossing occurs if successive samples have different algebraic signs. The zero-crossing rate is a measure of frequency content in the signal. Unvoiced speech exhibits a higher zero crossing rate than voiced speech or silence. The sampling frequency of the speech signal also determines the time resolution of the zero-crossing measurements. The zero-crossing rate corresponding to the i-th segment of the speech is computed as (3):

$$ZCR_i = \sum_{n=0}^{N-1} \left| \mathrm{sgn}[x_i(n)] - \mathrm{sgn}[x_i(n-1)] \right| \tag{3}$$

Where N=240, corresponding to 30 ms, denotes the length of the speech segment, xi(n).

3.1.3. Product of ZCR and Average Energy: For both of ZCR and average energy are considered simultaneously, product of them are calculated as (4):

$$A_i = E_i \times ZCR_i \tag{4}$$

Ai is always becoming lower at the syllable boundary than in the syllable.

3.1.4. Ratio of ZCR to Energy: There is another parameter, ratio of Ei to ZCRi, should be calculated to considering both of ZCR and average energy simultaneously, which is calculated as (5):

$$B_i = E_i / ZCR_i \tag{5}$$

Bi of unvoiced segment is generally much lower than that of voiced segments.

3.1.5. Ratio of Low Frequency Average Energy to total Average Energy: Each speech segment is decomposed into four different bands using a 3-level dyadic DWT, and the average energy of each band is computed.

In general, an unvoiced speech segment should show energy concentration in the high frequency bands, while a voiced segment should show energy concentration in its fundamental frequency bands of the wavelet domain. Because the fundamental frequency of voiced segments is ranged from 50-500Hz, the ratio of 500Hz below energy to total energy is computed and used in the method as the last parameter in voiced/unvoiced judging.

Let EH, be the high frequency (500Hz above) energy of a speech segment and EL be the low frequency(500Hz below) energy of a speech segment. Let Ej, be the energy in wavelet band –j. We can compute EH and EL as (6), (7):

$$E_H = \sum_{j=1}^{3} E_j \tag{6}$$

$$E_L = E_4 \tag{7}$$

Ratio of 500Hz below energy to total energy can be computed as (8):

$$R_i = E_L / \left(E_H + E_L \right) \tag{8}$$

So Ri represents the ratio of 500Hz below energy to total energy of the i-th segment of the speech.

3.2. Syllable Segmentation Based on the Features

Figure 2, shows the flow of the syllable segmentation Based on the features.

The pre-filtered speech segments should be deal with in sequence from the very beginning. For each segment, the computed features should be used to determined its type.

There are three types a segment should be one of them. The three types is non-speech, unvoiced, voiced. The non-speech segments are mostly like silence. The unvoiced segments are corresponding to the Mandarin consonants and the voiced segments are corresponding to the Mandarin vowels. But here is a exception. One category of consonants, the sonorants, is regarded as voiced speech, because its features computed are very similar to the vowels.

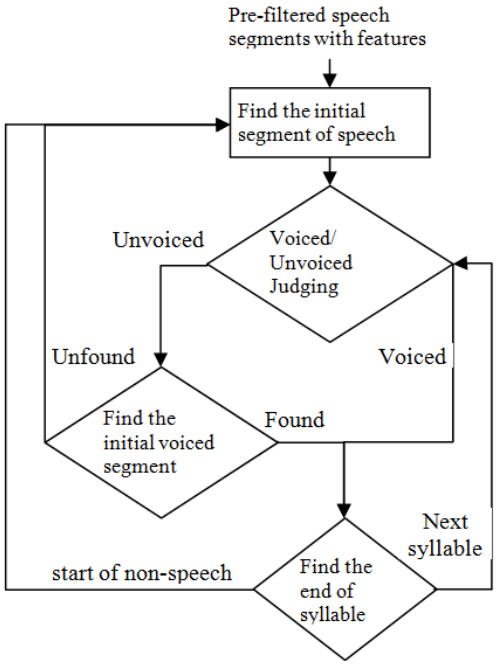

Figure 3. Flow Diagram of Syllable Segmentation

In Figure 3, We can see there are four step in the method flow.

3.2.1. Find the Initial Segment of Speech: Feature Ei is used in this step. The threshold of Ei is computed by four times of the 50ms signal at the very beginning which is always non-speech. If Ei is bigger than the threshold, the corresponding segment should be found as the initial segment of speech and then we should judge whether it is voiced or unvoiced.

3.2.2. Voiced/Unvoiced Judging: Features ZCRi, Bi, Ri is used in this step. The threshold of Ri is 0.2 and 0.8. If Ri is bigger than 0.8, the segment is voiced. If Ri is less than 0.2, the segment is unvoiced.

If Ri is between 0.2 and 0.8, the segments nearby should be judged by Ri first. Then compute the threshold of ZCRi and Bi by the nearby segments which have been determined voiced/unvoiced. ZCRv and Bv is average value of the voiced segments and ZCRu and Bu is average value of the unvoiced segments. Then, we can get the threshold ZCRt and Bt as (9), (10):

$$ZCR_t = \left(ZCR_u + ZCR_v\right)/2 \tag{9}$$

$$B_t = \left(B_u + B_v\right)/2 \tag{10}$$

If ZCRi is less than ZCRt and Bi is bigger than Bt ,the segment is voiced. Otherwise, the segment is unvoiced.

3.2.3. Find the Initial Voiced Segment: In this step, voiced/unvoiced judging is also needed. For each next segment, we judge its type of voiced/unvoiced just like step 2 to find the initial voiced segment.

3.2.4. Find the End of Syllable: Features Ei, Ai is used in this step. If Ei is less than its threshold computed in step 1, the segment is non-speech and go to step 1. If Ei and Ai both reach a local minimum in nearby several segment and the changing range is reach 50%, then the segment is the transition period and next syllable should start at next segment. Otherwise, the segment is voiced and go to test the next segment.

By the steps are repeated in turn according to the flow diagram above, each segment should be classified into three type, non-speech, unvoiced and voiced and the syllable boundaries are also determined. So the boundaries of syllables and the boundaries between consonants and vowels have been determined clearly.

4. Pitch Detection

The segmented speech segments should be decomposed into seven different bands using a 6-level dyadic DWT. After that, local maxima of the wavelet coefficients should be found out by several rules and finally the pitch period result can be achieved.

The flow chart of this approach is shown in Figure 4.

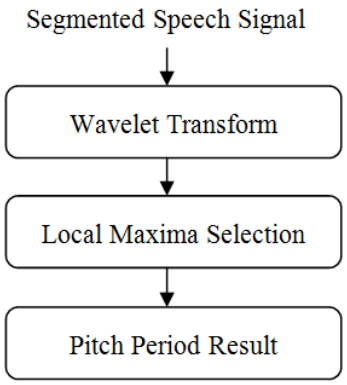

Figure 4. The Flow Chart of Pitch Detection Method

4.1.1. Wavelet Transform: The characteristics of wavelet transform and wavelet coefficients depend on wavelet function to be used. According to the abrupt change of speech signal at the closure of glottis, we should use smooth function to define the wavelet function. Finally, the quadratic spline wavelet with property of derivative is chosen to define the wavelet function in this paper. The corresponding filter coefficients of spline wavelet are listed in table 1.

Table 1. The Filter Coefficients of Quadratic Spline Wavelet

n	***-1***	***0***	***1***	***2***
h(n)	0.25	0.75	0.75	0.25
g(n)	0	1	–1	0

The best scales of wavelet decomposition for pitch detection can be determined by,

$$j = \mathrm{int}\left(\log_2\left(f_s / f_0\right)\right) \tag{11}$$

Where fs is the sampling frequency of the signal, f0 is the upper bound of the fundamental frequency. Because the fundamental frequency of the voiced speech signal is ranged from 60Hz to 500Hz, the best scales should be 4, 5, 6 when the input speech signal is sampled at 8 kHz.

So the input speech signal is decomposed into seven different bands using a 6-level dyadic DWT. Here, the wavelet coefficients of scale 4 correspond to frequency bands between 500Hz to 250Hz, the wavelet coefficients of scale 5 correspond to frequency bands between 250Hz to 125Hz and the wavelet coefficients of scale 6 correspond to frequency bands between 125Hz to 62.5Hz.

By experiments, we find using the wavelet coefficients of scale 5 performs better than using coefficients of the other two scales for most speech signals. So the wavelet coefficients of scale 5 is been chosen for pitch detection first.

If the pitch detection result is not consistent, the scale should be changed to 4 or 6 according to the average frequency of the result and do pitch detection again. Here, if the average frequency of the result is more than 250Hz, the scale should be changed to 4; otherwise, it should be changed to 6.

The pitch detection steps using wavelet coefficients in details are present below.

4.1.2. Local Maxima Finding and Selection: To estimate pitch period, we should find out the local maxima of the wavelet coefficients.

First, we find all local maxima in every 2ms. Because the upper bound of fundamental frequency of voiced speech signal is 500Hz which corresponds to 2ms.

So we can get pitch periods result roughly by those maxima. However, the result often has many errors. After analysis of the errors, we find the errors are mostly caused by several cases below:

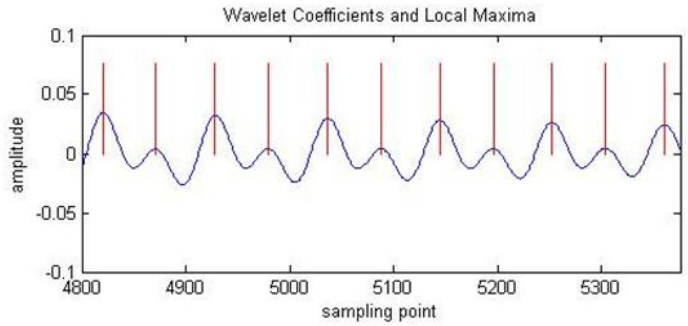

Figure 5. Error Case 1

In Figure 5, we can see about half of the local maxima should not be used as bound of the pitch period and we can find that the amplitudes of these local maxima are much lower than the others. Therefore, the first step of maxima selection is to delete these lower ones.

Selection step 1: Delete the local maximum if its amplitude is less than 0.7 times of its previous one and is also less than 0.7 times of its next one.

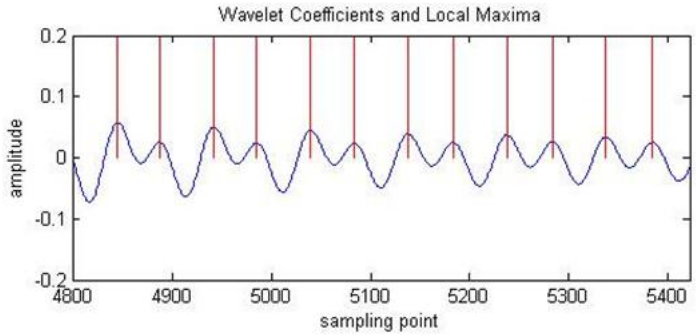

Figure 6. Error Case 2

In Figure 6, we can also see about half of the local maxima should not be used as bound of the pitch period and not all of these local maxima can be deleted in selection step 1 because its amplitude is much closer to the previous one. Here, we use another rule to judge which local maximum should be deleted. The rule is based on the minimum between two local maxima which we defined as local minimum. In Figure 3, we can see local minima before the local maxima which should be deleted are much higher than the other local minima. Therefore, we can get the second step of maxima selection as below,

Selection step 2: Delete the local maximum if absolute value of the local minimum before it is less than 0.5 times absolute value of the previous local minimum and is also less than 0.5 times absolute value of the next one.

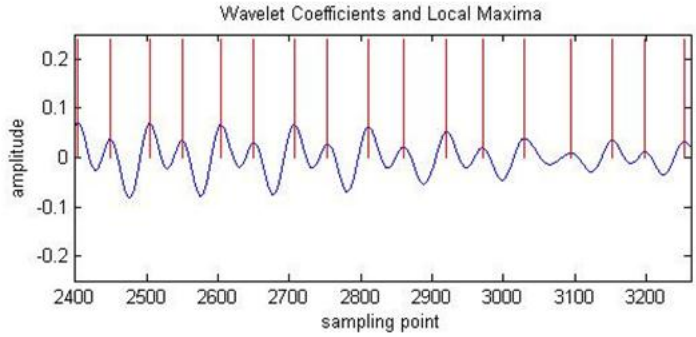

Figure 7. Error Case 3

In Figure 7, we can see the last several pitch periods is not much regular like the pitch periods in the front. There are some maxima should be deleted but it can't be picked out only by comparison of the local maxima and local minima in step 1 and 2. So we give another rule to judge which local maxima should be deleted in this case. We define the amplitude difference between the local maximum and the local minimum before it as local increasing difference (LIF), and define the amplitude difference between the local maximum and the local minimum after it as local decreasing difference (LDF).

Selection step 3: Delete the local maximum if its LIF is less than 0.9 times of the previous LDF and its LDF is less than 0.9 times of the next LIF.

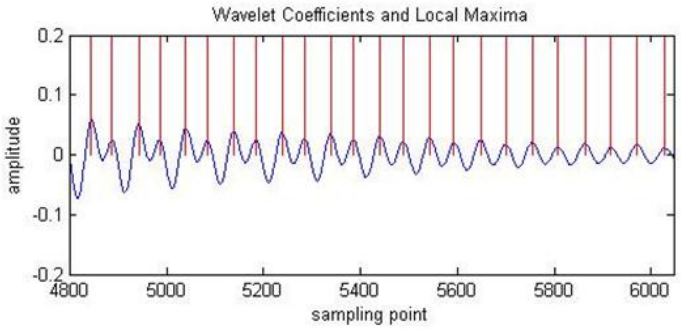

Figure 8. Error Case 4

In Figure 8, we can see the last few pitch periods is really difficult to judge which maxima should be deleted even by LDF and LIF. But if we reserve all maxima, the half pitch period errors will be made. Here, we find the maxima in the front pitch period can be selected by step 1-3, thus according to gradual changing of pitch period, we can delete maxima based as the front pitch period method.

Selection step 4: Delete the local maximum if its second previous maximum is deleted and the length of result pitch period is close to the previous one.

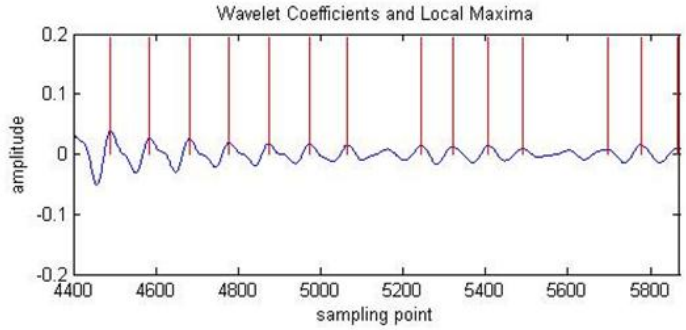

Figure 9. Error Case 5

In Figure 9, we can see there are several pitch periods is not much regular like the other pitch periods and the maxima in the periods may be deleted by step 1-3. But if these maxima are deleted, the double pitch period errors will be made. In order to avoid these errors, we give another rule for call back these deleted maxima. Also we can find the amplitudes of the maxima that should be called back are very small in the speech signal.

Selection step 5: If the delete local maximum causes its pitch period length is about twice than its previous pitch period and its amplitude is very small (less than 0.05 times of the max value of the speech signal), this maximum should not be deleted and we shall call it back.

After these five steps of selection, the local maxima left can be used to compute pitch periods. The distance between two successive local maxima is considered as the pitch period.

5. The Experimental Results

In this section, we evaluate the results of the proposed pitch detection method. The tests were performed using a large database comprising a wide variety of speech records, for different speakers and utterances. Three female and two male speakers were recorded reading Chinese words and sentences. The signals were ranging from 2-10s in length, sampled at 8 kHz, and organized into 100 speech files corresponding to a total about 4000s time length of speech signal. The database includes reference files containing pitch period mark. The proposed method was tested in varying noise conditions. White Gaussian noise of different intensity was added.

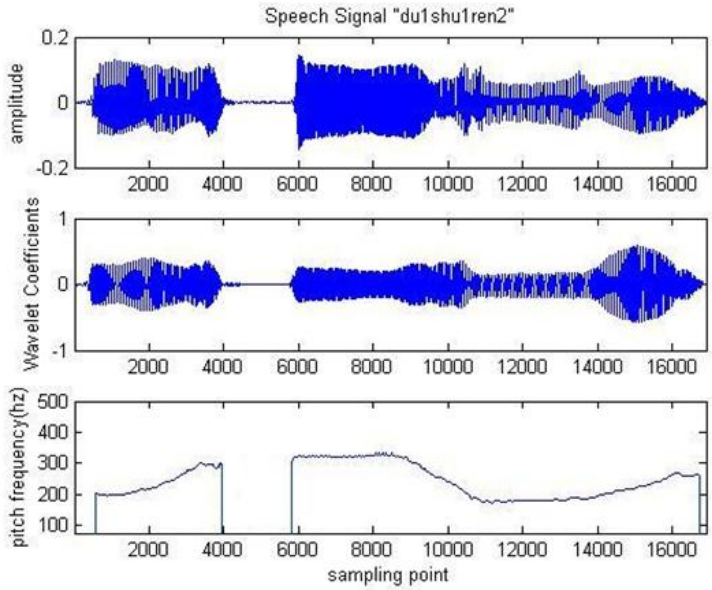

Figure 10. An Example of Pitch Detection

Figure 10 shows the performance of the proposed method to the speech signal of a three Chinese words 'du2shu1ren2'. We can see the pitch frequency contour is consistent and the result is satisfactory.

Table 2. Performance of Different Methods

SNR	Pitch Period Error Rate	
	The Proposed Method	***Weighted Wavelet method***
Clean	0.19%	0.95%
20	0.61%	1.22%
10	1.27%	3.60%
5	2.69%	5.78%

Table 2 shows the pitch period error rate[9] of the basic wavelet method and the proposed approach. Obviously, the performance of the proposed algorithm is superior to the weighted wavelet method in both clean and noise environment.

6. Conclusions

In this paper, we have given a pitch detection method which include pre-filtering, syllable segmentation and pitch detection. Syllable segmentation can make the following pitch detection more accurate. Maxima selection method of wavelet coefficients is used for pitch detection. It has been shown that the proposed pitch detection method based on pre-filtering, syllable segmentation and local maxima selection exhibits superior performance compared to weighted wavelet method in both clean and noise environment.

Acknowledgements

This work has been supported by Tianjin Science and Technology Development Foundation of High School (20090805) and Research Foundation of Tianjin University of Science and Technology (20100204).

References

[1] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection", IEEE Transactions on Acoustics Speech and Signal Processing, vol. 25, **(1977)**, pp. 24-33.
[2] M. J. Ross, H. L. Shaffer and A. Cohen, "Average magnitude difference function pitch extractor", IEEE Transactions on Acoustics Speech Signal Processing, vol. 22, **(1974)**, pp. 353-362.
[3] A. M. Noll, "Cepstrum pitch determination, "The Journal of the Acoustical Society of America", vol. 41, **(1967)**, pp. 293-309.
[4] S. Kadame and G. F. Broudreaux-Bartels, "Application of wavelet transform for pitch detection", IEEE Trans on IT, vol. 38, **(1992)**, pp. 917-924.
[5] D. Charalampidis and V. B. Kura, "Novel Wavelet-based Pitch Estimation and Segmentation of Non-stationary Speech", 2005 8th International Conference on Information Fusion, vol. 7, **(2005)**, pp. 1-5.
[6] A. N. Akansu and R. A. Haddad, "Multiresolution Signal Decomposition: transforms subbands wavelets", San Diego: Academic Press, **(2001)**.
[7] C. Bao, "Principles of digital speech coding", Xi'an: Xidian University Publishing House, **(2007)**.
[8] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Englewood Cliffs, **(1978)**.
[9] L. R. Rabiner, M. J. Cheng and A. E. Rosenberg, "A comparative performance study of several pitch detection algorithms", IEEE Trans on Acoustics, Speech and Signal Processing, vol. 24, **(1976)**, pp. 399-418.

Authors

Runshen Cai

Runshen Cai received his B.E degree from College of Marine Geosciences, Ocean University of China, China. After that he obtained his B.E and PhD degrees from College of Information Technical Science, Nankai University, China. He is currently working as a lecturer in Computer Science and Information Engineering College, Tianjin University of Science and Technology, China