The acoustic features based on pitch detection after process analysis

Ying Ma[1a], Chao Chen[1] ,Maoshen Jia[2] ,Shanji chen[1]

[1]School of Physics and Electronic Information Engineering,Qinghai University for Nationalities, 810000,China

[2] Beijing industry university,100000,China

[a]email: 1037125248@qq.com

Keywords: Speech signal; Cepstrum method; Autocorrelation method; Pitch period;

Abstract. In the field of speech signal processing, a pitch detection algorithm (PDA) is commonly used to estimate pitch or fundamental frequency. If the given speech signal is clean, the algorithm can achieve better detection result. However, in general, the speech signal would inevitably influenced by the background noise, the detection algorithm may not work well and the detected pitches may deviated from the correct position. In this paper we propose a new approach for improving the accuracy of pitch detection. This approach uses a median filter to remove the outliers in the results produced by short-time autocorrelation or cepstrum method. The conducted experiments show that the proposed approach works well.

Introduction

In the field of digital speech signal processing, the accuracy of extracted parameter from the speech signal is very important for many speech processing algorithms. Only accurate values of parameters represent the essential features of a speech signal, these values can be used for the high quality speech synthesis, speech recognition and voice compression. Pitch detection is of particular importance in speech signal processing, since the result of extraction directly affects whether the synthetic speech can accurately restore the original speech signal spectrum [2].

The estimation of pitch period is called Pitch Detection, and the ultimate goal of pitch detection is to draw the pitch trajectory curves that are entirely consistent with the vibratory frequency of the vocal folds. If it is difficult to obtain the real trajectory, a trajectory curve as consistent as possible should be found, which could be applied for speech compression coding, speech synthesis and speech recognition, etc.

The most common used pitch detection algorithms, such as the short-time autocorrelation and the cepstrum method, may produce pitch detection errors due to the background noises. These errors lead one or several pitch estimations to deviate from the normal values (usually deviated from the normal values by twice or 1/2) in pitch track. This kind of eccentric point is known as the "outlier" of the pitch track [4]. These outliers are not the fundamental frequency and should be removed. In order to remove these outliers, a variety of smoothing algorithms can be used. The most common used algorithms are median smoothing algorithm and linear smoothing algorithm. In this paper we propose a method for improving pitch detection. The method is consisted of two steps: firstly the given speech signal is processed by the short-time autocorrelation or the cepstrum method; secondly the initial detected result is smoothed to remove the outliers using a median smoothing strategy. Finally, we show the simulated results to validate the effectiveness of the proposed method.

The Cepstrum algorithm

It is known that a speech signal is not an additive signal, but a convolutional signal. In order to process the speech signal through a linear system, it is processed firstly using the convolutional homomorphy system. The output of the convolutional homomorphy system is a pseudo-temporal sequence called the "complex cepstrum" of the original sequence. The definition of complex cepstrum is expressed [1] as:

$$\hat{x}(n) = IFT\{\ln[FT\{x(n)\}]\} \qquad (1)$$

The corresponding cepstrum is expressed [1] as:

$$c(n) = IFT\{\ln|FT[x(n)]|\} \qquad (2)$$

The main difference between cepstrum and complex cepstrum is that the cepstrum is the inverse Fourier Transform of log-spectral amplitude of sequence *x(n)* and it is the even symmetrical component of complex cepstrum. Both of them convert the convolutional operation into addition operation in the pseudo time-domain. Therefore, the signal can be processed by a linear system having superposition. Moreover, the complex cepstrum involves logarithmic operations of complex numbers, while the cepstrum just involves logarithmic operations of real numbers. It is obvious that cepstrum needs much less number of operations than that of complex cepstrum and becomes faster.

On the other hand, spectrum analysis has the following merits: the waveform in time domain is subject to change caused by the exterior environment, while the spectrum of the speech signal has a good robustness to these kind of the changes. Moreover, since the spectrum of speech signal owns strong acoustic characteristics, it is very significant to extract speech features such as Mel-Frequency Cepstrum Coefficients (MFCC), formant, and pitch period by applying frequency domain analysis.

The short-time autocorrelation

The short-time autocorrelation function is defined as:

$$R_n(\tau) = \sum_{m=0}^{N-1-\tau} [s(n+m)w(m)][s(n+m+\tau)w(m+\tau)] \qquad (3)$$

where $\{s(n)\}$ is a speech signal with energy limited; τ is a shift distance , while $w(m)$ is the even symmetrical window function.

As for the short-time autocorrelation function method in pitch detection, the main principle is as follows: The estimation of pitch period is conducted by comparing the similarity between the original signal and its displacement signal. If the shift distance is equal to the pitch period, then two signals have the largest similarity.

Although short-time autocorrelation function owns a strong ability to suppress noises, it is prone to be affected by half pitch error and double pitch error.

Pitch detection algorithm based on short-time autocorrelation function is a common method. It is known that this algorithm is especially suitable for pitch extraction in noise environment. Short-time autocorrelation function appears to be a peak value in the pitch period, the interval between two adjacent peaks is a pitch period. Generally, the fundamental component is not always the strongest component. Rich harmonic component makes speech signal waveform become very complex. All of these make it difficulty to detect pitch and therefore the errors such as half pitch error and double pitch error are produced. Other hardness is that there are also cases of voiced and voiceless in the signal, which lead pitch detection become problematic.

The post-processing of pitch period detection

To remove the outliers in pitch detect, we use a median filter to smooth the initial detected result. Suppose *x(n)* is a signal consisted of a series of pitches detected by the cepstrum method or by short-time autocorrelation method, *y(n)* is the output of median filter. An sliding window is applied onto the input signal, the window is centered at *n*. Set the number of samples be *L* on both sides of *n*. So we have *2L+1* samples in the sliding window. These samples are sorted according to their magnitudes, the value come to the middle is used as the output of smoothing filter. Empirically, *L* can be set as 1 or 2 such that a sliding window can cover 3 or 5 samples. Median smoothing has a merit that it not only remove a bit of outliers but also preserve step respond between two smoothed segments in the pitch trajectory [4].

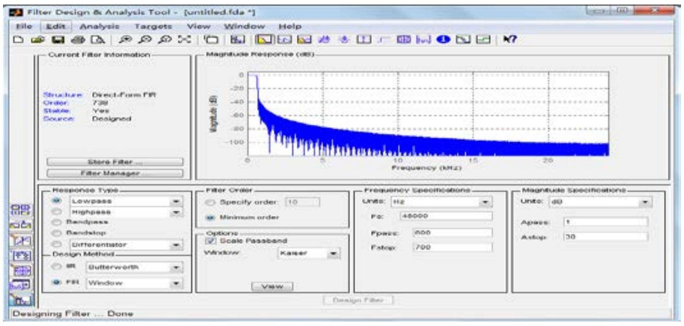
The implementation and discussion

To verify the proposed method, we conducted an experiment. We use a sequence of speech signal recorded using Cooledit system. The recording is proceeded in a common indoor environment. The sound source if from a Tibetan male and the sampling rate of the speech signal is 44kHz with single channel. The speech signal is used for estimating pitch period and then the outliers are removed.

Speech Signal Preprocessing

To enhance the high frequency resolution, the input speech signal had been pre-processed by weighing the high frequency parts and removing the effects from lip radiation[4].

In our experiment, we used the FDATool (Filter Design & Analysis Tool) of MATLAB, which can implement various filter designs, analysis and performance evaluation. The procedure of pre-processing is as follows: typing FDAtool in MATLAB command window, the FDATool page out the needed information; Entering the parameters of filter, the MATLAB statements are generated as follows.



Current Filter Information

Structure: Direct-Form FIR
Order: 738
Stable: Yes
Source: Designed

Response Type

● Lowpass
○ Highpass
○ Bandpass
○ Bandstop
○ Differentiator

Design Method

○ IIR Butterworth
● FIR Window

Filter Order

○ Specify order: 10
● Minimum order

Options

☑ Scale Passband
Window: Kaiser

Frequency Specifications

Units: Hz
Fs: 48000
Fpass: 600
Fstop: 700

Magnitude Specifications

Units: dB
Apass: 1
Astop: 30

Designing Filter ... Done

Figure 1 Filter designing panel of FDATOOL

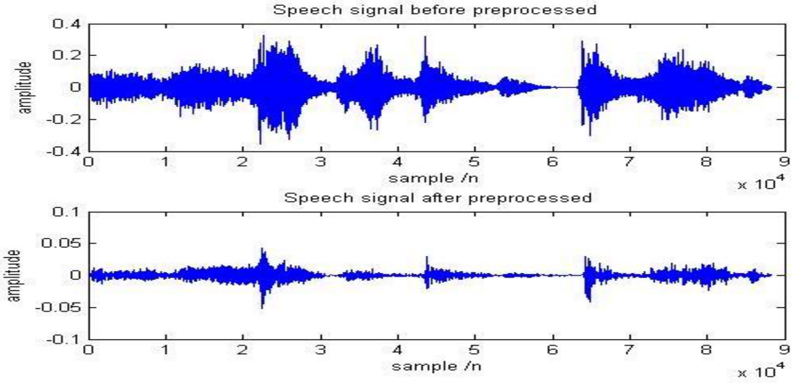


Figure 2 Speech signal after preprocessed

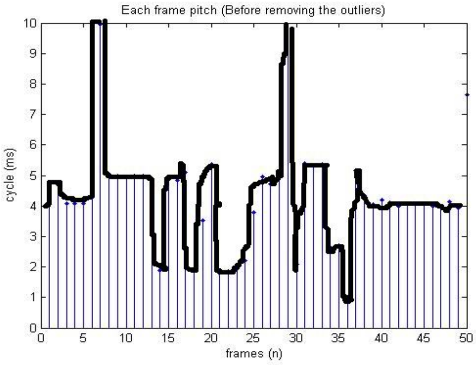
Figure 1 shows the filter designing panel of FDATools and figure 2 illustrates the waveforms of speech signal before and after preprocessing. Since the amplitude of signal with high frequency was enhanced, the variation of spectrum is not too sharp. Therefore it is more effective to extract feature parameters from speech signal.

Post-processing of speech signal pitch

The above pre-processed speech signal is further processed using short-time autocorrelation method and cepstrum method to detect pitches. In the case of short-time autocorrelation, there are several outliers in the detected pitch periods as shown in figure 3(a). The outliers are removed by using median filter mentioned in Section 4 and the result is shown in figure 3(b) which is clean and accurate by visual observation. When cepstrum method was used to estimate pitch periods of the pre-processed speech signal, there existed many outliers as illustrated in figure 4(a) compared with that of short-time autocorrelation method. As the same, a median filter is adapted to this initially detected result signal, the final result is shown in figure 4(b). The outliers are almost removed completely.

We found in the simulation that satisfied result can be obtained when the width of window is at least larger than double of pitch period. The largest pitch period is about 20*ms*, so the width of window should be larger than 40*ms*. We set the length of frame be 40*ms* with 1764 samples. 50

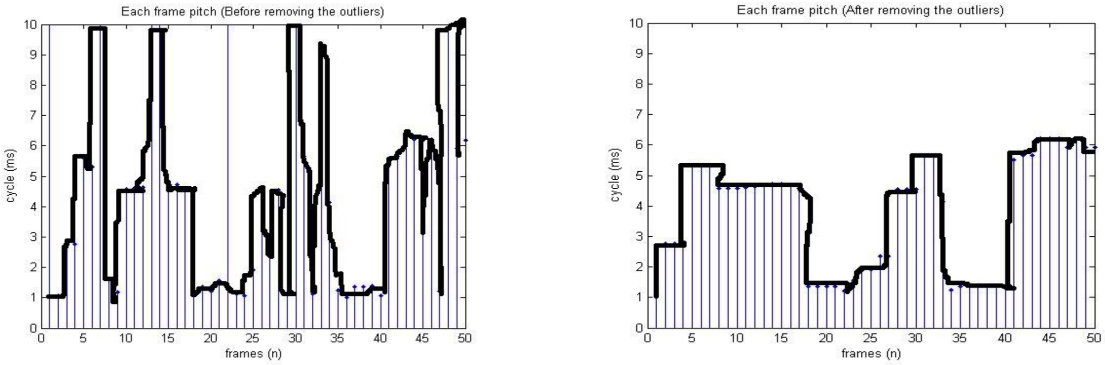
frames of speech signal was used for extracting pitches by applying short-time autocorrelation method and cepstrum method. The extracted pitch periods are relatively correct.



(a) Short-time autocorrelation analysis removing the outliers)

(b) Short-time autocorrelation analysis (Before (After removing the outliers)

Figure 3 Short-time autocorrelation analysis



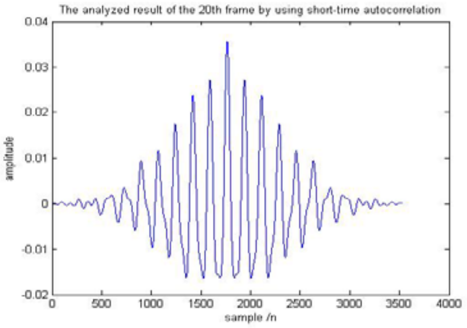
(a) Cepstrum analysis
(Before removing the outliers)

(b) Cepstrum analysis
(After removing the outliers)

Figure 4 Cepstrum analysis

By observing figure 3(a) and figure 4(a), the results of these two pitch detection methods have great differences before removing the outliers. This will result in that the estimated periods of one or several pitches deviates from normal positions in the pitch period trajectory. These deviated pitched are outliers, which lead to an erroneous judgment during pitch detection. When the outliers are removed as shown in figure 3(b) and figure 4(b), the differences of the above two detection algorithms greatly decreased. Generally, the differences are between 5*ms*~7*ms* for the used 50 frames. Since the detected pitch in the 20th frame deviated from the normal trajectory, we used this frame with sample points from 85281 to 85281+1764 for further analysis in our experiment.

The analyzed results of 20th frame by using short-time autocorrelation analysis and cepstrum analysis are illustrated in Figure 5 and Figure 6, respectively.



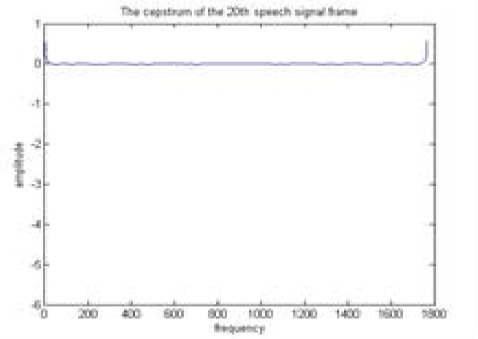


Figure 5 Analyzed result of the 20th frame by using short-time autocorrelation

Figure 6 Cepstrum of the 20th speech signal frame

The response of autocorrelation function applied to consonant sound is periodic in speech signal. The function will reach its maximum after sampling with a fixed interval. The pitch period for a consonant sound can be estimated by finding the distance between the maximum point and the next peak point of autocorrelation function. Since the fundamental frequency is generally in the range from 40Hz and 1000Hz, the possible incorrect points in front should be removed to avoid erroneous judgment. However, when computing pitch period, the number of removed sample points should be compensated, that is, the pitch period T = (the distance between maximum point and the next peak point adds the number of removed sample points) divides sampling rate.

By observing the figure 5, we know that the pitch period of speech signal is

$$T=(1938-1765+40)/44000=4.84ms$$

The detected result is on the trajectory of pitch period, which has little error and consistent with the result when using 50 frames speech signal.

Figure 6 shows the result by using cepstrum method applied to the 20th frame of speech signal. Since the cepstrum is almost a line in the horizontal direction, it is difficult to read the maximum. In fact, the position of the maximum is at 93 when using MATLAB “max” function. As for cepstrum method, the pitch period of consonant sound can be estimated by using the inverse of maximum peak position, that is,

$$T=1 / 93=10.7ms$$

This detected result deviates from the trajectory of pitch period, and has large error compared with the result by using 50 frames.

We repeated the above experiment and found that deviation occurs when using both of these two algorithm. The deviation of cepstrum is large than autocorrelation in most case.

The detected pitch periods have large difference before and after removing outliers for both methods. The averaged period after removing outliers is $T=(4.84 + 10.7) / 2 = 7.77ms$, which is almost consistent with that most pitch periods are in the range $5ms$~$7ms$ as shown in figure 4.

Conclusion

When extracting pitch period from a speech signal, the errors may occur. The detected pitches deviate from normal trajectory. We proposed a method for improving the estimation of pitch period from speech signal. The method removes outliers using a median filter applied to the initially detected results produced by short-time autocorrelation method and cepstrum method. However, this is not effective for any frame, it may fail for some frames we found in our simulation. Since no pitch detection method can reach enough accuracy, in practice, a solution is detecting pitch periods by using multiple algorithms and then averaging the detected pitch period to obtain relatively correct results. Developing more accurate detecting algorithm is under our consideration for future work.

Acknowledgement

In this paper, Qinghai University for Nationalities Science and Technology Cooperation Fund No. 312000514315008 and Qinghai Province Science and Technology Department of Basic Research Program No. 2015-2J-721 and Project of Qinghai-International Science and Technology under Grand No. 2014-HZ-821 .

References

[1] Zhao Li. Speech signal processing – 2nd edition [M]. Beijing: mechanical industry press, 2009.5.

[2] Gao Xiquan, Ding Yumei. Digital signal processing [M]. Xi 'an: xi 'an university of electronic science and technology press, 2008.8.

[3] Zhang Zhichong. Proficient in MATLAB R2011a [M]. Beijing University of aeronautics and astronautics press, 2011.11.

[4] Xue-ying zhang. Digital speech processing and MATLAB simulation [M]. Electronic industry press, 2010.07.

[5] Rob. A power law detector based on autocorrelation function [J]. Applied acoustics. 2010.09:352

[6] Sundarrajan Rangachari,Philipos C.Loizou.A Noise-Estimation Algorithm for Highly Non-Stationary Environments,Speech Communication,2006,48(2):220