AN EFFECTIVE PITCH DETECTION METHOD FOR SPEECH SIGNALS WITH LOW SIGNAL-TO-NOISE RATIO

ZHEN-DONG ZHAO[1], XI-MEI HU[1], JING-FENG TIAN[2]

[1]Department of Electronic and Communication Engineering, North China Electric Power University, Baoding, 071002, China
[2]North China Electric Power University Science & Technology College, Baoding 071002, China
E-MAIL: huximei_ncepu@yahoo.cn, tianjfhxm_ncepu@yahoo.cn

Abstract:

Pitch detection in noisy environment plays an important role in speech analyzing and recognition. In this paper, an effective pitch detection method is proposed. Noised speech is denoised by an improved form of spectral subtraction method. A linear predictive coding analysis is performed on the segmented speech, and the segmented speech is filtered by the inverse filter to give the linear prediction error. The cepstrum of the linear prediction error and the autocorrelation function of the cepstrum are calculated. The result of the simulation shows that compared with the autocorrelation function pitch detection method, a distinct improvement in effect can be seen by using this improved method in pitch detection.

Keywords:

Pitch detection; spectral subtraction; linear prediction error; cepstrum; autocorrelation function

1. Introduction

The pitch period is an important parameter in the analysis and synthesis of speech signals, so the pitch detection is an essential component in a variety of speech processing systems. Besides providing valuable insights into the nature of the excitation source for speech production, the pitch contour of an utterance is useful for speaker identification/verification and for speech instruction to the hearing impaired and voice disease diagnostics[1].

At present, a wide variety of sophisticated pitch detection methods have been proposed in the speech processing literature.

1) Waveform-based estimating method. It uses the speech waveform to estimate the pitch period, such as Data reduction method (DARD) and Parallel processing method (PPROC)[1].

2) Correlativity processing method. It is comparatively insensitive to phase distortion and is very simple in the disposal of hardware structure, including Autocorrelation method using short-term clipping (AUTOC) and Average magnitude difference function (AMDF)[1] .

3) Transforming method. It transforms speech signal from time domain to frequency or inverse spectrum domain to estimate period, for example, Cepstrum method (CEP) and Wavelet method which combine the time and frequency domain features[1].

Because there are a lot of problems in pitch detection[2], no one algorithm has been developed so far performing perfectly for all different speakers, applications and environmental conditions.

The autocorrelation function is suitable to the pitch detection in noisy environment. But usually the fundamental frequency is quite close with the first formant frequency, and using the autocorrelation function method only often causes half frequency error and double frequency error in pitch detection[3]. Cepstral method was one of the conventional methods in pitch detection. The result could be accurate when clean voiced speech was processed. However, error would increase greatly when noisy speech or the edge of voiced speech is processed[4]. In view of this kind of question, an effective pitch detection method is proposed in this paper: Firstly, noisy speech is denoised by an improved form of spectral subtraction method, and then the cepstrum of the linear prediction error and the autocorrelation function of the cepstrum are calculated.

2. Pitch detection algorithm

2.1. The improved spectrum subtraction[5]

The spectrum subtraction is one kind of effective speech enhancement technology. The basic diagram of spectrum subtraction is as follows in figure 1. In the figure, $s(n)$ is the clean speech, $d(n)$ is the Additive White Gaussian Noise, $\lambda_n(k)$ is the coefficient of the power

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

spectrum of noise. Y_k (k=0, 1···) is the frequency spectrum coefficient of the noised speech *y(n)*, and S_k (k=0, 1···) is the frequency spectrum coefficient of the clean speech $s(n)$.

$$x(n) = s(n) + d(n) \qquad (1)$$

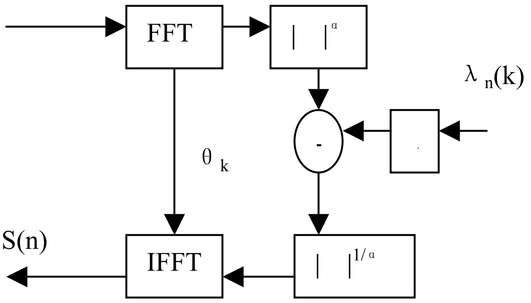


Figure 1.

The spectrum amplitude coefficient $\left|\hat{S}_k\right|$ of the enhanced speech $\hat{s}(n)$ is obtained by the equation:

$$\left|\hat{s}_k\right| = \left[\left|Y_k\right|^{\alpha} - \beta\lambda_n^{\alpha}(k)\right]^{\frac{1}{\alpha}} \qquad (2)$$

where α and β (β>=1) are two parameters. When α=2, β=1, this algorithm is the traditional spectrum subtraction. If α and β are adjusted to appropriate value, a better enhanced effect can be obtained. The experiment shows that when α =2, β =5 the enhanced effect surpasses the traditional spectrum subtraction greatly .This algorithm is the improved form of spectrum subtraction.

2.2. Algorithmic analysis

1) The interaction between the vocal tract and the glottal excitation disturbs the detection from glottal excitation, so we use linear prediction error to eliminate the vocal tract information which can improve the accurate to some extent.

The speech signal during LPC analyses can be described as follows[1]:

$$s(n) = \sum_{k=1}^{p} \alpha_k s(n-k) + Gu(n) \qquad (3)$$

where α_k denotes the prediction coefficient, *p* is the prediction rank number, *u(n)* is glottal excitation and *G* denotes the range gene. The prediction error filter can be described as:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \qquad (4)$$

We use LPC method to analyze the speech and get the prediction coefficient α_k which makes up of the prediction error filter *A(z)*, and then we let speech signal pass the prediction error filter and the prediction error *e(n)* (equals to *Gu(n)*) is obtained.

Regards to voiced speech, the prediction error *e(n)* is stronger in each start place of pitch period, so it can be used in pitch detection. Because the spectrum of prediction error approaches smoothly, when we use it in pitch detection, the influence of formant peaks can be reduced, such as half frequency and double frequency.

2) The cepstrum can be described as follows:

$$c(n) = IFT\left\{\ln\left|FFT\left[e(n)\right]\right|\right\} \qquad (5)$$

which has a strong peak corresponding to the pitch period of the voiced-speech segment being analyzed[1].

3) The short-term autocorrelation function of *x(n)* is defined as[6]:

$$R_w(l) = \sum_{n=-\infty}^{\infty} x_w(n) x_w(n+l) = \sum_{n=0}^{N-l-1} x_w(n) x_w(n+l) \qquad (6)$$

It has the property as follows[6]: If the sequence *x(n)* has period N_p, the autocorrelation function of *x(n)* is a period function and has the same period N_p as *x(n)*, that is, if $x(n) = x(n+N_p)$, then $R(k) = R(k+N_p)$.

The surd has no period. The autocorrelation function of surd is not a period function, and *R(k)* will be attenuated quickly with the increasing of *k*. The voiced-speech has period and its autocorrelation function has the same period as it.

4) In virtue of the above analysis, an effective pitch detection method is proposed: Firstly, noisy speech is denoised by an improved form of spectral subtraction method, and then the cepstrum of the linear prediction error and the autocorrelation function of the cepstrum are calculated.

2.3. Post-processing

In view of different wrong points in the fundamental frequency, which has been calculated with the above method, we could use the search tentative smooth algorithm[7].

If $f_1, f_2, \cdots, f_N$ denote continuous *N* frames of fundamental frequency and $f_i^{'}$ is the smoothed fundamental frequency of the i^{th} frame, when f_i is processed, the half frequency and double frequency is processed first. The processing method is as follows:

If $\left|\frac{f_i}{2} - f_{i-1}\right| < c_1$, then makes $f_i^{'} = \frac{f_i}{2}$;

If $|2f_i - f_{i-1}| < c_1$, then makes $f_i' = 2f_i$.

The double frequency and half frequency situation is processed, and then we process the stochastic wrong situation.

If $|f_i - f_{i-1}| > c_1$ and $|f_{i+1} - f_{i-1}| > c_2$, then $f_i' = 2f_{i-1} - f_{i-2}$;

If $|f_i - f_{i-1}| > c_1$ and $|f_{i+1} - f_{i-1}| \le c_2$, then $f_i' = \frac{(f_{i-1} + f_{i+1})}{2}$.

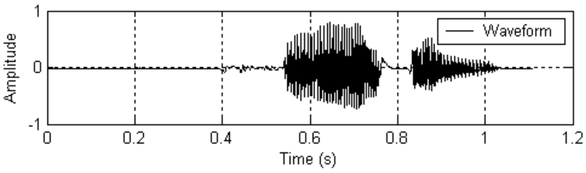
If the above conditions can not be fulfilled, then makes $f_i' = f_i$.

c_1, c_2 are the threshold which is decided by the experiment. c_1 is the threshold of the frequency difference between two continuous frames, and that c_2 is the threshold of the frequency difference between two continuous frames which are at intervals of one frame.

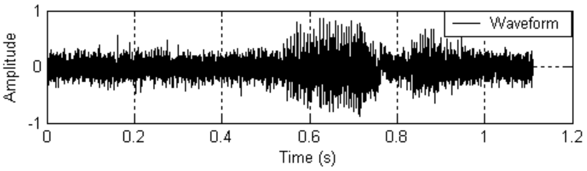
Because the fundamental frequency usually changes continuously and slowly, according to the sampling frequency and the frequency range pronounced by persons, usually c_1 is 10 and c_2 is 25. The function of the two thresholds and above rules is to limit the variety of fundamental frequency between neighboring frame not to surpass c_1, and that the variety of fundamental frequency between neighboring frame which are at intervals of one frame not to surpass c_2. Therefore, the fundamental frequency trace can be smoothed.

3. Simulation experiment

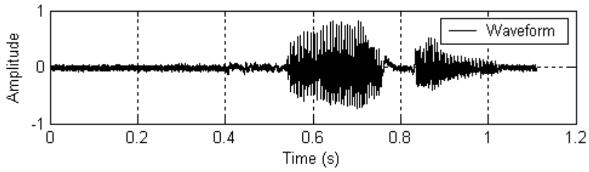
In the following simulation experiments, the male utterance 'Hua Bei' is used with sampling frequency of 8 kHz. The noisy speech signals are generated by adding White Gaussian Noise in different SNR to it. The original speech signal is shown in Figure 2.a. The noisy speech signal generated by adding White Gaussian Noise in 0dB to it is shown in Figure 2.b. Figure 2.c is the speech which is denoised by spectrum subtraction. The traditional autocorrelation function method and the method proposed in this paper are compared in pitch detection for the above sentence. The results are shown in Figure 3 and Figure 4.



(a) The clean speech

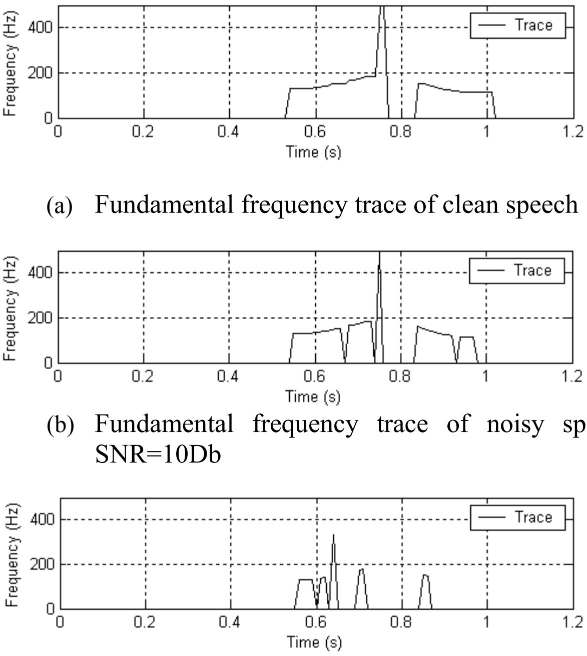


(b) The noisy speech, SNR=0dB



(c) The speech denoised by improved spectrum subtraction

Figure 2.

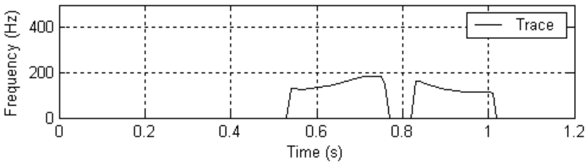


(a) Fundamental frequency trace of clean speech

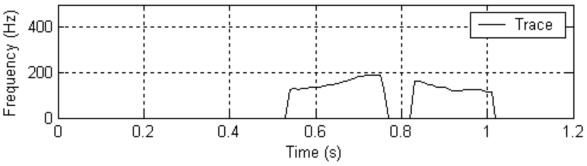
(b) Fundamental frequency trace of noisy speech, SNR=10Db

(c) Fundamental frequency trace of noisy speech, SNR=0dB

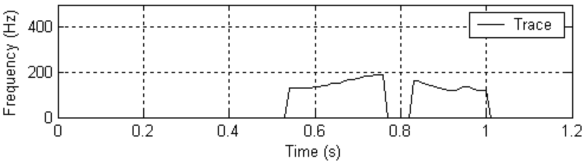
Figure 3. The experiment results of conventional autocorrelation function



(a) Fundamental frequency trace of clean speech



(b) Fundamental frequency trace of noisy speech, SNR=10dB



(d) Fundamental frequency trace of noisy speech, SNR=0dB

Figure 4. The experiment results of the method proposed in this paper

From Figure 3 and Figure 4, we can see that compared with the conventional pitch detection method, a distinct improvement in effect can be seen by using this improved method in pitch detection. This improved method behaves more robustly than the conventional method, especially effective at low signal to noise ratio.

4. Conclusions

In this paper, an effective pitch detection method is proposed: Firstly, noised speech is denoised by the improved form of spectral subtraction method, and then the cepstrum of the linear prediction error and the autocorrelation function of the cepstrum are calculated. It overcomes the disadvantage of the autocorrelation function—the half and double frequency error often happens in the low SNR environments. It also overcomes the stochastic error problem when the speech signals have a comparatively big change curve.

The result of the simulation shows that compared with the conventional pitch detection method, a distinct improvement in effect can be seen by using this improved method in pitch detection.

References

[1] Hui Ding, Bo Qian, Yanping Li, and Zhenmin Tang, “A Method Combining LPC-Based Cepstrum and Harmonic Product Spectrum for Pitch Detection Article Title”, Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 537-540, December 2006.

[2] T. Shimamura, and H.Takagi, “Noise-Robust Fundamental Frequency Extraction Method Based on Exponentiated Band-Limited Amplitude Spectrum”, the 47th IEEE International Midwest Symposium on Circuits and Systems, pp. 141-144, 2004.

[3] Jing Bai, and Gang Wei, “A New Method for Speech Signals Pitch Detection Based on LPC and Autocorrelation”, Master’s thesis, unpublished, South China University of Technology, Guangzhou, China, 2005.

[4] Xaoya Wang, “The Application of Cepstrum in Pitch and Formant Extraction”, Radio Engineering, China, vol. 34, no. 1, pp. 57-58, 2004.

[5] Yifang Xu, Jinjie Zhang, Kaisheng Yao, “Speech Enhancement Applied to Speech Recognition in Noisy Environments”, Journal of Tsinghua University (Science and Technology), China, vol. 44, no. 1, pp. 41-44, 2001.

[6] Li Zhao, Speech Signal Processing, China Machine Press, Beijing, China, 2003.

[7] Ying Hu, Ning Chen, and Xu Xia, “Pitch Detection Using a Improved Algorithm Based on ACF”, Electronic Science and Technology, China, no. 2, pp. 5-28, 2007.