Pitch Detection Using Variable Bandwidth LPF

가변 대역폭 LPF를 이용한 피치 검출

Hong KEUM*, Guemran BAEK*, Myungjin BAE*, Ho Sung Jang**

금 홍*, 백 금 란*, 배 명 진*, 장 호 성**

ABSTRACT

In speech signal processing, it is very important to detect the pitch exactly. Although various methods for detecting the pitch of speech signals have been developed, it is difficult to exactly extract the pitch for wide range of speakers and various utterances. Thus we propose a new pitch detection algorithm which takes advantage of the G-peak extraction. It is a method to detect the pitch period of the voiced signals by finding MZCI(maximum zero-crossing interval) of the G-peak which is defined as cut-off bandwidth rate of LPF(low pass filter). This algorithm performs robustly with a gross error rate of 3.63% even in 0 dB SNR environment. The gross error rate for clean speech is only 0.18%. Also it is able to process all courses with high speed.

요 약

음성신호 처리에서, 피치를 정확하게 찾아내는 것이 매우 중요하다. 현재까지 많은 피치 검출 방법들이 제안되어 왔지만, 광범위한 화자와 다양한 음성 데이터로부터 정확한 피치를 찾는 것은 어렵다. 따라서 본 논문에서는 G-peak 검출을 이용한 새로운 피치 검출 알고리즘을 제안한다. 이 방법은 G-peak의 MZCI(최대 영교차 간격)을 LPF(low-pass filter)의 차단대역폭으로 결정하여 음성신호의 피치를 검출하는 방법이다. 본 알고리즘은 0dB SNR 환경 하에서 3.36%의 그로스 에러를 나타내는 잡음에 강인한 방법이다. 또한, 잡음이 없는 음성의 그로스 에러는 0.18%였고, 모든 과정은 고속 처리가 가능하다.

I. INTRODUCTION

Determining the pitch period of a speech waveform is an important step in pitch-synchronous analysis of short-term quasi-stationary periodic data [1][10][11]. In the analysis, we can use the pitch to obtain proper vocal tract parameters. We can use the pitch to easily change, to maintain the naturalness and intelligibility of quality in speech synthesis. Also we can use the pitch to eliminate the personality for speaker-independence in speech recognition.

A lot of methods for the pitch detection have been proposed until now. The pitch detection algorithms can be categorized as the methods in time domain, in frequency domain, and in time-frequency hybrid domain. The methods in time domain generally emphasize the periodicity of voiced speech before detecting the pitch by using a decision logic. These algorithms are based on parallel processing, average magnitude difference function(AMDF), autocorrelation, harmonics matching, etc.[2][6]. Since these methods do not

*숭실대학교 정보통신공학과
**홍익대학교 전자전산기공학과
접수일자 : 1994년 8월 1일

need to perform a transformation into any domain, the computation time to find the pitch can be reduced. Also the detected pitch period exhibits a good resolution due to detecting the pitch in time domain. However, these methods may bring about the errors, when there are some phonemic transitions within the analysis frame and the speech signals are corrupted by background noises.

In frequency domain, the pitch period is usually measured by the spectral intervals between the harmonics of speech spectrum. Generally, the spectrums are based on a frame ; e.g., 20~40msec length. Because the effects of phonemic transitions and background noises averaged in this frame, the effects for extracting the pitch lessened. However, when one wants higher frequency resolution, the computation time required to process these methods must be taken longer to increase the number of FFT points. The pitch detection algorithms in frequency domain are the methods of harmonics detection, lifter banks, comb-filtering, etc.[3].

The last method for pitch detection is to process in time-frequency hybrid domain. These methods take some good characteristics in both time and frequency domain. There are the methods of analysing cepstrum, comparing with the spectrum, etc.[5]. One of problems for these methods is a lot of computations due to transform time(or frequency) domain into frequency(or time) domain.

Although various methods for detecting the pitch of speech signals have been developed, it is difficult to exactly extract the pitch from the wide range of speakers and the various utterances.

Accordingly, in this paper, we propose a new pitch detection algorithm that gives a good perfo·rmance and resolves the processing complexity. After performing the variable bandwidth LPF, we detected the pitch in the G-peak waveform[6, 9]. In section Ⅱ, we briefly review the production model of voiced speech signals. In section Ⅲ, we define the G-peak and propose our algorithm to extract the pitch by using variable bandwidth LPF. Finally, some computer simulations are given in section Ⅳ.

Ⅱ. SPEECH PRODUCTION MODEL

In the speech production model, the excitation source of unvoiced speech signals is a random noise generator. The unvoiced speech has no periodicity and higher average zero-crossing rate than the voiced signal, because it has the first formant with wide bandwidth at near 3 kHz.

On the other side, generally, the excitation source of voiced speech is a glottal pulse train that has quasi-periodic pulse and large amplitude. The voiced speech signals have periodicity owing to vibrating of vocal tract. Due to the resonance of vocal tract, the voiced speech has formant with bandwidth. Therefore, the voiced waveform has damped-oscillation in a pitch period. In frequency domain, the spectrum of voiced speech is represented as multiplication between the harmonics of fundamental frequency and the formant envelope of vocal tract. Since the gain of the first formant(F_1) is generally higher 10dB than that of the remaining formants, the resonance of vocal tract can be approximated by envelope of only F_1.

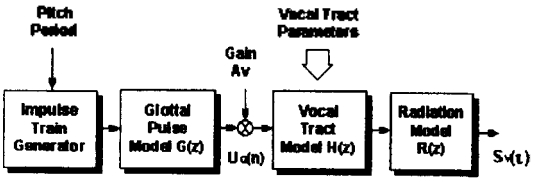

Fig. 2-1 Speech production model for voiced signals

The envelope of the first formant in frequency domain can be approximated as a cosine form. In time domain, the waveform may be obtained through inverse Fourier transform(suppose that the phase is zero) as follows ;

$$h(t)=\int_{-\infty}^{\infty} F(f)e^{j2\pi ft}\,df$$

$$\int_{-B_W/2}^{B_W/2}\cos\left(\frac{2\pi f}{2B_W}\right)e^{j2\pi ft}\,df=2\cos(2\pi F_1 t)\int_0^{\frac{B_W}{2}}$$

$$=\frac{4B_W}{\pi-4\pi B_W^2}\cdot\frac{1}{t^2}\cos(\pi B_W t)\cos\left(2\pi F_1 t-\frac{\pi}{2}\right),\quad (2\text{-}1)$$

The glottal pulse shape can be modeled as the following equation by Rosenberg[6] :

$$g(n)=\begin{cases}\frac{1}{2}\left[1-\cos\left(\frac{\pi n}{N_1}\right)\right], & 0\le n\le N_1\\ \cos\left[\frac{\pi(n-N_1)}{2N_2}\right], & N_1\le n\le N_1+N_2\\ 0, \textit{ otherwise.} & \end{cases}\quad (2\text{-}2)$$

Thus, the speech signal, s(n), is roughly approached with Eq. (2-1) and Eq.(2-2) in time domain.

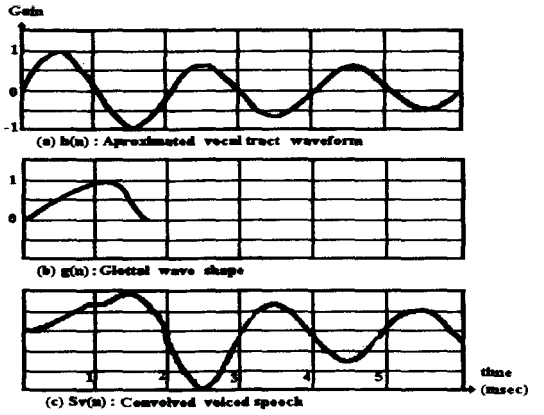

Fig. 2-2 The approximation analysis for voiced speech.
(a) h(n) : impulse response of the approximated vocal tract.
(b) g(n) : glottal waveform.
(c) s(n) : voiced speech waveform by h(n)*g(n).

Fig. 2-2 shows an example waveform of Eq. (2-1), Eq. (2-2), and Eq. (2-3), respectively. The first positive peak of the waveform in a pitch period of voiced signal is especially distinguished from the other peaks. That is shown in Fig. 2-2 (c). The reasons are that the first formant, F_1, is damped-oscillation in a pitch period and the glottal pulse is asymmetric for the zero level. That is, the G peak is defined as the peak that is mainly affected by the glottal pulse characteristics in a pitch interval. Conclusively, we can define the first peak as the G-peak and do remaining as side-peaks.

Ⅲ. PITCH EXTRACTION USING THE G-PEAK

The G-peak is defined as the first peak of voiced signal and it is obtained from the convolution of glottal waveform and vocal tract waveform in time domain. The zero-crossing interval(ZCI) of the G-peak in voiced speech is longer than that of side-peaks. Since the first formant has some bandwidth, the waveform of voiced speech has damped-oscillation in a pitch period. Thus, the magnitude of the G-peak is larger than that of side-peaks.

Because the speech signal is convolved with many formants and glottal-pulses, it is very difficult to detect only the G-peak in the voiced speech waveform. Also, the formants and the G-peak of speech signals are time-variant. Therefore, before detecting the G-peak for voiced speech, it is desirable to remove the higher formant of speech signal. To do this, the voiced speech is passed by the low-pass filter as the following equation.

$$s'\left(n-\frac{N}{2}\right)=\sum_{i=0}^{N-1}s(n-1),\quad (3\text{-}1)$$

where N is a bandwidth interval of the filter, because cutoff frequency, f_T, relates to $f_T=f_S/N$ (or $N=f_S/f_T$). To adaptively reject an effect of formant in the G-peak detection, the cut-off frequency of LPF, f_T, must be varied in each frame. Resultingly, in this paper, we take cut-off frequency of the filter by using the properties of the G-peak. Because the ZCI of the G-peak in a pitch interval is the longest one, the detected maximum ZCI becomes interval of the G-peak.

Before finding the maximum ZCI, we must take the zero-crossing point, Zc(i). Then, ZCI(i) is to subtract Zc(i) from Zc(i+1) as follows :

$$ZCI(i) = Z_c(i+1) - Z_c(i), \quad (i=0, 1, 2, 3, \ldots). \quad (3\text{-}2)$$

Where $Z_c(i)$ stands for the *i-th* zero-crossing point and Zc(i+1) for the *(i+1)-th*. The bandwidth interval of the LPF is roughly estimated by the maximum ZCI as follows :

$$N \Rightarrow Max\{ZCI(0), ZCI(1), \cdots, ZCI(M-1)\}, \quad (3\text{-}3)$$

where M is the number of zero-crossing points of the waveform in a frame.

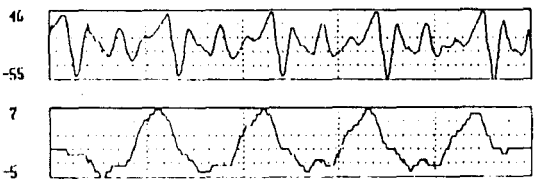

Fig. 3-1 G-peak detection using second-order variable bandwidth LPF.
(a) Speech signals
(b) The waveform through second-order varible bandwidth LPF

We process Eq. (3-1) two times with the resultant value, N. This indicates that the voiced signal is processed by second-order LPF. Therefore, the G-peak in a pitch period may be properly distinguished from side peaks such as Fig. 3-1(b). Since s'(i) is asymmetrical for ground, to remove side-peaks, the threshold level for the G-peak can be taken by the maximum value of side-peaks. The decision logic is presented as the following equation in speech signal,

$$\text{Pitch} = \frac{N_E - N_S}{PZCIR}, \quad (3\text{-}4)$$

where we define N_S as the starting point of the first detected G-peak and N_E as one of the last detected the G-peak in the analysis frame.

According to Eq.(3-4), we can find ZCP(zero crossing point) of voiced signals that is processed by second-order variable bandwidth LPF. After N_S and N_E are determined in that waveform, the interval between both points is obtained. Therefore, it is the pitch that is the interval between N_S and N_E divided by positive zero crossing interval rate (PZCIR) in a frame.

Ⅳ. EXPERIMENT AND RESULT

For computer simulation, we use the IBM-PC/486 DX2(50) interfaced with A/D converter. The speech signal was sampled at 8kHz, lowpass filtered and digitized with a 16bit A/D converter. Four sentences pronounced five times by three males and two females speakers were used for simulation.

Utterance 1) "INSUNE KOMANUN CHUNJAESONYUNWL JOAHANDA"
Utterance 2) "JESUNIMKESEO CHUNJICHANGJOWI KIOHUNWL MALSUMHASEOSSDA"
Utterance 3) "SOONGSILDAE JUNGBOTONGSINKONGHAKWA UMSEONGSINHOCHURIYUNGUSIL'"
Utterance 4) "GONG IL RI SAM SA O YUK CHIL PAL GU SIP"

The experimental procedure is shown in Fig. 4-1. In analysis, the length of one frame is 256 samples and each adjacent frame is overlapped by 128 samples.

In frequency domain, the vocal tract resonance is multiplied by the fundamental frequency. Voiced signals are divided into the vocal tract and vocal cord. Both elements are convolved in time domain, then the first envelope will be eminent. That is, we obtain the G-peak which is influenced by glottis. In this experiment we find ZCP, ZCI and MZCI in each frame and settle N with MZCI.

We use second-order LPF with variable cut-off bandwidth, N. Finally, the pitch is obtained by using the G-peak and decision logic. Fig. 4-2 represents the pitch contour about speech signals. This figure shows the prominent reduction of halving, doubling, and tripling error. Also we obtain smoothing pitch contour such as Fig. 4-2(b).

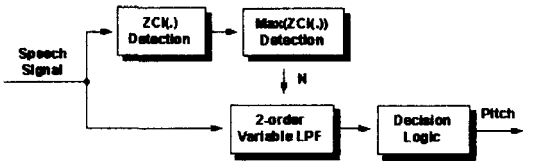

Fig. 4-1 Block diagram on pitch detection

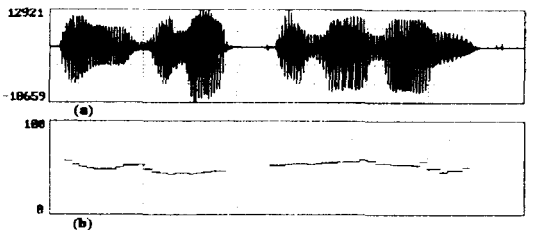

Fig. 4-2 Pitch contour
(a) Speech signal
(b) Pitch contour

Table 1 represents the gross error rates for each speech sample. The gross error rate is defined as follows : we compare the result of our algorithm with the eye-checked result. When the result of our algorithm differs with the eye-checked pitch by more than 1 msec for a frame, we increase the error count by 1. This 1 msec corresponds to 8 samples. If there are 7 frames that contain errors, the gross error rate in that case would be

$(7/62) * 100 = 11(\%)$.

As can be showned in Table 1, this experimental result gives robust performance with a gross error rate of 3.63% even in 0 dB SNR environment.

Table 1. The gross error rates for each speech sample.

Utterances	no. of analyzed frames	gross error rates(%) clean speech	SNR 6dB	SNR 3dB	SNR 0dB
1	192	0.00	1.04	1.04	3.64
2	192	0.00	0.52	1.04	3.12
3	192	0.52	1.04	1.04	3.12
4	64	0.00	0.00	0.00	1.56
average	630	0.18	0.91	1.09	3.63

The gross error rate for clean speech is only 0.18%. We did not consider the fine error, because a time difference is less than 1 msec. Since there were virtually none, fine errors occur when the pitch detector allows a poor resolution to reduce computation time, or when the resolution in the transform domain is low.

V. CONCLUSION

We have developed a novel algorithm that determines the pitch period of speech in real time. Pitch extraction is one of the most important part in speech processing. If we obtain the pitch accurately, then the pitch can be used in the analysis of the vocal tract parameter without the influences of vocal cord. It can be used to maintain the naturalness and intelligibility in speech synthesis and also to obtain high accuracy of speech recognition because of reducing the influences by speaker.

In this paper we proposed the new algorithm about pitch detection by using variable bandwidth LPF. The algorithm uses the G-peak which is found by LPF. The bandwidth of LPF must be varied, because the bandwidth of the G-peak and the formant rate are varied at each frame. Thus, we have to apply the variable bandwidth LPF. That is, we ought to apply the variable cut-off bandwidth rate to LPF in order to emphasize the G-peak and decrease the formant effects. As

above-mentioned, the pitch can be detected.

We wish that you refer to the detailed procedure in the previous section, and our argument is supported by the experimental results. Owing to this algorithm, we improved the accuracy of pitch detection and extracted it with a high speed. As appears can be shown in Table 1, the experimental results give robust performance with a gross error rate of 3.63% even in 0 dB SNR environment.

REFERENCES

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech signals*, Englewood Cliffs, Prentice-Hall, New Jersey, 1978.
2. P. E. Papamichalis, *Practical Speech Processing*, Prentice-Hall, Inc. . Englewood Cliffs, New Jersey, 1987.
3. S. Seneff, "Real Time Harmonic Pitch Detection," *IEEE Trans Acoust. Speech and Signal Processing*, vol. ASSP-26, pp. 358-365, Aug. 1978.
4. S. D. Stearns and R. A. David, *Signal Processing Algorithms*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.
5. M. Bae, and S. Ann, "Fundamental Frequency Estimation of Noise Corrupted Speech Signals Using the Spectrum Comparison," *J Acoust., Soc., Korea*, vol. 5, no. 3, pp. -, 1989.
6. E. Lee, C. Park, M. Bae, and S. Ann "The High speed Pitch Extraction of Speech Signals Using the Area Comparison Method," *The Korean Institute of Telematics and Electronics*, vol. 22, no. 2, pp. 101-105, 1985.
7. M. Bae, J. Rheem, and S. Ann "A Study on Energy Using G-peak from the Speech Production Model," *The Korean Institute of Telematics and Electronics*, vol. 24, no. 3, pp. 381-386, 1987.
8. Hans Werner Strube, "Determination of the instant of glottal closure from the speech wave," *J Acoust. Soc. Am.*, vol. 5, no. 5, pp. 1625-1629, Nov. 1974.
9. M. Bae, I. Chung, and S. Ann, "The Extraction of Nasal Sound Using G-peak in Continued Speech," *The Korean Institute of Telematics and Electronics*, vol. 24, no. 2, pp. 274-279, 1987.
10. W. Hess, *Pitch Determination of Speech Signals*, New York : Springer- Verlag, 1983.
11. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Trans. on Acous., Speech and Signal Processing*, vol. ASSP-24, pp. 399-417, Oct. 1976.

▲Hong Keum : Vol.13, No.1E 참고

▲Guemran Baek : 현재 한국과학기술원 석사과정

▲Myungjin Bae : Vol.13, No.1E 참고

▲Ho Sung Jang : Vol.12, No.3 참고