Pitch Detection Using Variable LPF

Hong KEUM, Guemran BAEK, Myungjin BAE

Department of Telecommunication Engineering
SoongSil University
1-1, Sangdo-5 dong, Dongjak-ku
Seoul 156-743
Republic of Korea

ABSTRACT In speech signal processing, it is very important to detect the pitch exactly. The algorithms for pitch extraction that have been proposed until now are not enough to detect the fine pitch in speech signal. Thus we propose the new algorithm which takes advantage of the G-peak extraction. It is the method to find MZCI(maximum zero-crossing interval) which is defined as cut-off bandwidth rate of LPF(low pass filter) and detect the pitch period of the voiced signals. 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 course with high speed.

I. INTRODUCTION

In speech signal processing, it is very important to extract the pitch exactly [1]. In the analysis, we can use the pitch to obtain properly the vocal tract parameter. It can be used to easily change or to maintain the naturalness and intelligibility of quality in speech synthesis and 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 emphasize generally the periodicity of the voiced speech before deciding on the pitch by using a decision logic. These algorithms are parallel processing, average magnitude difference function(AMDF), autocorrelation, harmonics matching, etc. [2,6]. Since these methods do not need to perform the transformation into any domains, the computation time to find the pitch can be reduced. Also the detected pitch period exhibits a good resolution due to deciding on the pitch in time domain. However, these methods may be brought 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 obtained on a frame basis; e.g., 20~40msec length. Because the effects of phonemic transitions and background noises are averaged for this frame, the effects for extracting the pitch are 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 the 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 cepstrum analysis, comparing with the spectrum, etc. [5]. One of problems for these methods is to have a lot of computations due to transform time domain from/to frequency domain.

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.

Accordingly, in this paper, we propose the new pitch detection algorithm that gives a good performance and resolves the processing complexity. After performing the variable bandwidth LPF, the pitch is detected by the G-peak [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 the random noise generator. The unvoiced speech has no periodicity and appears higher average zero-crossing rate than the voiced signal, because it has the first formant with wide bandwidth at near 3 kHz. 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 formants with bandwidth. Therefore, the voiced waveforms in a pitch period have damped-oscillation. In frequency domain, the spectrum of voiced speech appears to be multiplied the harmonics of fundamental frequency by formant envelope of vocal tract. Since the gain of the first formant(F_1) is generally higher 10dB than that of the remain formants, the resonance of the vocal tract can be approximated by envelope of only F_1.

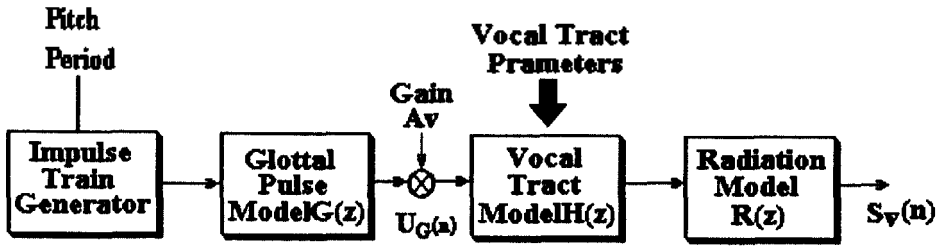

Fig. 2-1. Speech production model for voiced signals

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

$$\begin{aligned} h(t) &= \int_{-\infty}^{\infty} F(f) e^{j2\pi ft} df \\ &= \int_{-Bw/2}^{Bw/2} \cos\left(\frac{2\pi f}{2B_W}\right) e^{j2\pi ft} df \, 2\cos\left[(2\pi F_1 t) - \frac{\pi}{2}\right] \\ &= \frac{4B_W}{\pi - 4\pi B_W^2} \frac{1}{t^2} \cos(\pi B_W t) \cos\left(2\pi F_1 t - \frac{\pi}{2}\right). \end{aligned} \tag{2-1}$$

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

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

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

$$s(n) \fallingdotseq h(n) * g(n) \tag{2-3}$$

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 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 remainings as side-peaks.

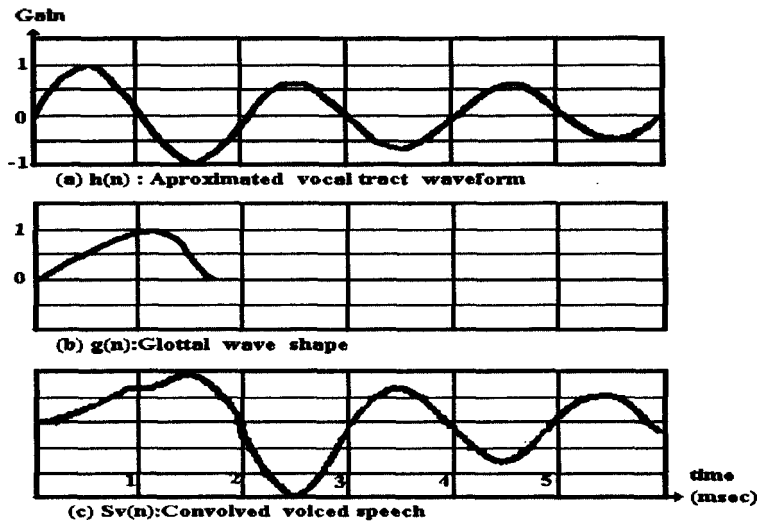

Fig. 2-2. The approximation analysis for voiced speech
(a) h(n) : impulse response of the approximated vocal tract
(b) g(n) : glottal waveform
(c) s(n) : voiced speech waveform by h(n)*g(n)

III. PITCH EXTRACTION USING THE G-PEAK

The G-peak is definded as the first peak of voiced signal that is obtained the convolution of glottal waveform and vocal tract waveform in time domain. The zero-crossing interval(ZCI) of 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 G-peak in the voiced speech waveforms. Also, the formants and G-peaks of speech signals are time-variant. Therefore, preceding to detect the G-peak for voiced speech, it is desirable to remove the higher formants of speech signal. To do that, the voiced speech is passed by the low-pass filter as following equation.

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

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 formants in G-peak detection, the cut-off frequency of LPF, f_T, must be varied in each

frame. Resultly, in this paper we take cut-off frequency of the filter by using the properties of G-peak. Because the ZCI of G-peak in a pitch period 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, $Z_c(i)$. Then, ZCI(i) is to subtract $Z_c(i)$ from $Z_c(i+1)$ as follows;

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

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

$$N \doteq \max \{Z_c(0), Z_c(1), \cdots, Z_c(M-1)\} \tag{3-3}$$

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

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

$$\text{Pitch} = \frac{N_E - N_S}{ZCIR} \tag{3-4}$$

According to Eq.(3-4), we find ZCP(zero crossing point) of voiced signals that is processed by second-order variable LPF. After starting(N_S) and ending

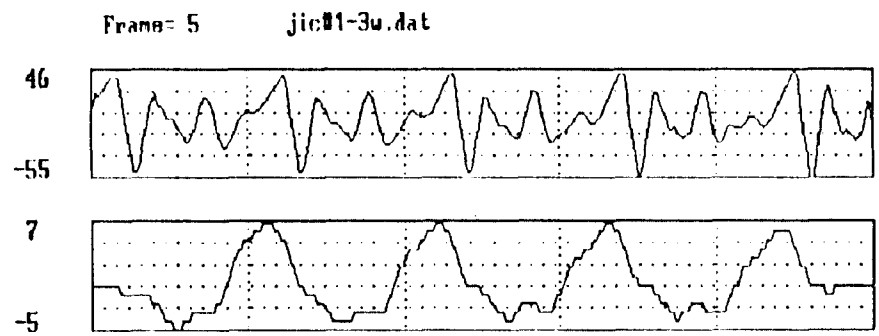

Fig. 3-1. G-peak detection using second-order variable LPF
(a) Speech signals
(b) The waveform through second-order variable LPF

point(N_E) are determined in that waveform, the interval between both points is obtained. Therefore, it is the pitch that the interval between N_S and N_E divided by zero crossing inerval rate(ZCIR) in a frame.

Ⅳ. COMPUTER SIMULATION

The speech signal was sampled at 8 kHz and lowpass filtered at 4 kHz and digitized with a 16 bit A/D converter. The speech data are comprised of 5 speakers. 3 males and 2 females. The following sentences were spoken.

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 process 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, then first envelope will be eminent. That is, we obtain G-peak which is influenced by glottis. In this experimentation 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. This figure shows the prominent reduction of halving, doubling, and tripling error. Also we obtain smoothing pitch contour such as Fig. 4-2.

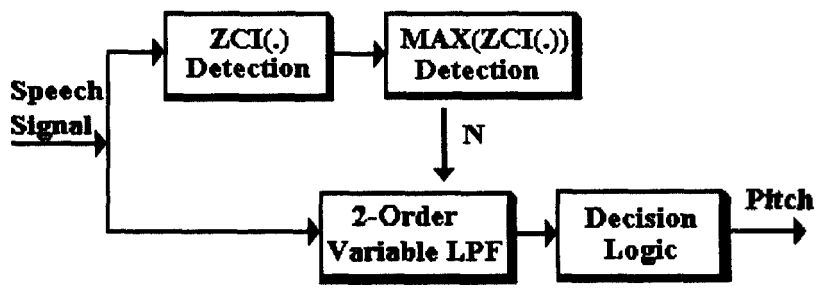

Fig 4-1. Block diagram on pitch detection

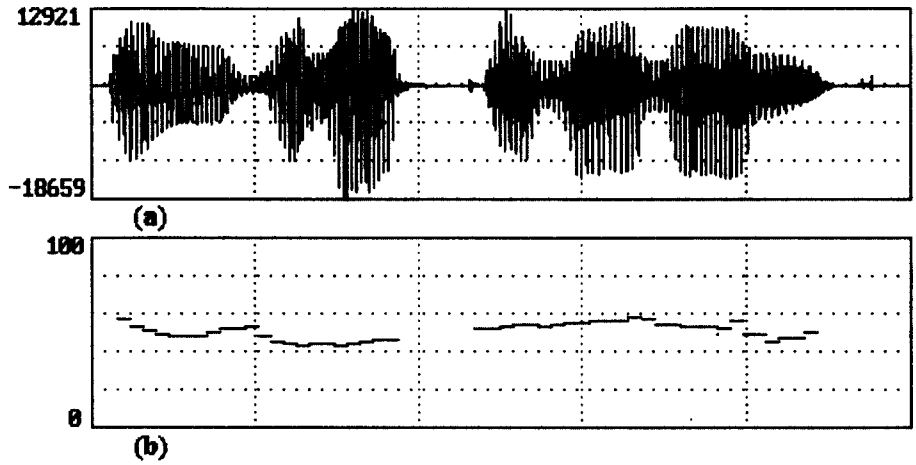

Fig. 4-2. Pitch contour
(a) Speech signal (b) Pitch contour

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 that of eye-checked. 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 in that case would

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

As appears Table. 1, this experiment result gives robust performance with a gross error rate of 3.63 % even in 0 dB SNR environment. The gross error for clean speech is only 0.18 %. We did not consider the fine error, because the error is less than 1 msec time difference. Since there were virtually none, fine errors occur when the pitch detector let the resolution become poor to reduce computation, or when the resolution in the transform domain is low.

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

V. CONCLUSION

We proposed the new algorithm about pitch detection. Pitch extraction is one of the most important problems in speech processing. If we take the pitches 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 to obtain high accuracy of speech recognition because of being reduced the influences by speaker.

In this paper we proposed the new algorithm about pitch detection by using variable LPF. It uses the extracted G-peak, which is found by LPF. This value must be varied, because the bandwidth of G-peak and the formant rate are varied at each frame. Thus, we have to apply the variable LPF. That is, we ought to apply the variable cut-off bandwidth rate in order to emphasize G-peak and decrease formant. From the fact that mentioned above, the pitch is detected.

Owing to this algorithm, we obtained the pitch, improved the accuracy of pitch detection and extracted it with the high speed.

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 & R.A. David, *Signal Processing Algorithms*, Prentice-Hall, Inc., Engle wood Cliffs, New Jersey, 1988.

5. M. Bae, and S. Ann, "Fundamental Frequency Estimation of Noise Corrupted Speech Signals Using the Spectrum Comparis on", J.Acoust., Sot., Korea, Vol.5, No. 3, 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, November 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.