Detection of Glottal Closure Instant using the property of G-peak

G-peak의 특성을 이용한 성문폐쇄시점 검출

Hong Keum*, Daesik Kim*, Myungjin Bae*, Youngil Kim**

금 홍*, 김 대 식*, 배 명 진*, 김 영 일**

※이 연구는 1993년도 한국과학재단 연구비 지원에 의한 결과임

ABSTRACT

It is important to exactly detect the GCI(Glottal Closure Instant) in the speech signal processing. A few methods to detect the GCI of voiced speech have been proposed, untill now. But these are difficult to detect the GCI for wide range of speakers and for various vowel signals. In this paper, we proposed a new method for GCI detection using the G peak. The speech waveforms are passed through the LPF of variable bandwidth. Then, the GCI's of voiced speech are detected by the G-peak based on the filtered signals. We compared the detected with the eye-checked GCI at the SNR of clean, 20dB, and 0dB. We took into account the range within 1ms between eye-checked and detected GCI. We obtained the result of the detection rate as 97.9% in the clean speech, 96.5% in 20dB SNR, and 94.8% in 0dB SNR, respectively.

요 약

음성신호의 처리에서 GCI를 정확하게 검출하는 것은 중요하다. 따라서 이에 대한 연구가 부분적으로 진행되어 왔다. 이러한 방법은 광범위한 화자와 다양한 단어에 대해 적합하지 못하기 때문에 우리는 G-peak를 사용하여 GCI를 검출하는 새로운 기법을 제안하였다. 우선 음성 신호 파형을 가변 저역 통과 여과기에 통과 시킨다. 여과된 신호를 사용하여 G-peak를 검출하고 이를 기준으로 GCI를 검출하게 된다. 제안된 방법으로 검출한 GCI와 눈으로 찾은 GCI의 차이가 1ms이내이면 고려의 대상으로 삼았다. 제안된 방법은 검출율이 각각 0dB SNR하에서 94%, 20dB SNR하에서 96.5%, 무잡음에서 97.9%를 나타내었다. 결론적으로 제안된 방법은 잡음 환경하에서도 우수한 수행결과를 보였다.

I. Introduction

It is important to exactly detect the GCI(Glottal Closure Instant) in the speech signal processing. Supposing that the accurate GCI is extracted in speech signal processing, it may be ap

*Soongsil University
**Osan College

접수일자 : 1993년 10월 20일

plied to analyze the speech waveform to the pitch synchronously, to recognize with removing the personality of speakers, and to synthesize with high quality because we know the phase information of vocal cord.

A few methods to detect the GCI of voiced speech have been proposed, until now. Some of them use almost prediction error for the GCI[3, 8]. One locates the GCI using prediction error, but the presence of a pulse at the input of the vocal tract is not always predictable. Algorithms with more complete consideration based on this assumption fall under the heading of autocovariance matrix determinant evaluation[5]. This method, however, does not work well for all vowel signals ; indeed, some vowels cause great difficulty in determining GCI's. The problem is due to many significant residual pulses occuring around the GCI from both the input pulse and the large prediction error. Otherwise, this method is adequate, but computationally expensive. An alternative method uses the occurrence of discontinuities in derivatives of the glottal airflow[5]. This algorithm works well for most clean vowel signals. However, for vowels which do not possess manifest discontinuities in the derivatives of glottal airflow, e.g., strong deceleration of the airflow of front and high vowels during the closing glottis, they fail. A drawback of this kind of method is the confusion the discontinuities introduce by the derivatives or by the noise excitation and contaminant noise. Thus, the restrictions of clean data and of certain vowel signals are imposed on its application.

Hence, it is difficult to detect the GCI for wide range of speakers. Thus, in this paper, we propose an algorithm to detect the GCI by using the G-peak based on the speech production model. In the following section, we will explain the speech production model and define the G-peak. In section Ⅲ, we will account for algorithm which proposed in this paper. Then in section Ⅳ and Ⅴ, we will present the experimental results and give the conclusion, respectively.

Ⅱ. Speech production model

In the speech production model, the excitation source of unvoiced speech signal is the random noise generator as shown in Fig. 2-1. 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 3KHz. 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 as shown in Fig. 2-2(b). 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.

The envelope of first formant in frequency domain can be approached a cosine form as shown in Fig. 2-3. In time domain, the waveform may be obtained by inverse fourier transform(supposed that the phase is zero) as follows :

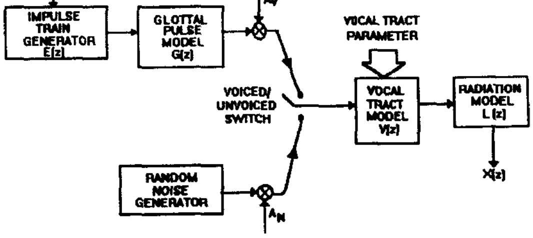


Fig. 2-1. The speech production model.[1]

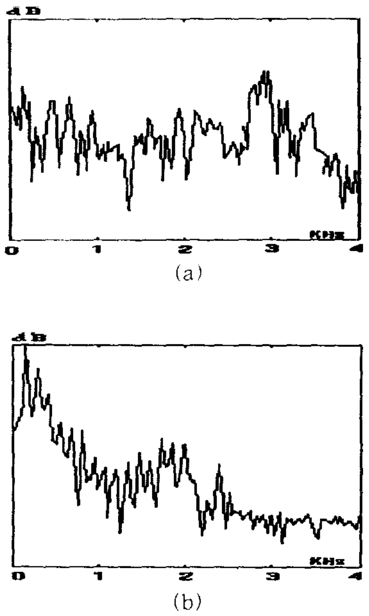


Fig. 2-2. The specturm of unvoiced and voiced speech.
(a) The spectum of unvoiced speech
(b) The spectum of voiced speech

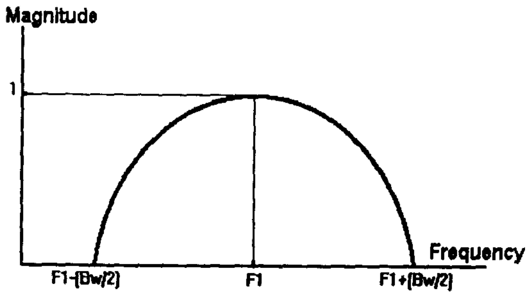


Fig. 2-3. An envelope approximation of the first formant in frequency domain.[5]

$$h(t)=\int_{-\infty}^{\infty} F(f)\, e^{j2\pi ft}\ df$$
$$=\int_{-Bw/2}^{Bw/2}\cos(\frac{2\pi f}{2B_W})\, e^{j2\pi ft}\, df\ 2\cos[(2\pi F_1 t)-\frac{\pi}{2}]$$
$$=\frac{4B_W}{\pi-4\pi B_W^2}\,\frac{1}{t^2}\,\cos(\pi B_W t)\,\cos(2\pi F_1 t-\frac{\pi}{2}). \qquad (2\text{-}1)$$

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

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

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

$$s(n) \doteqdot h(n) * g(n) \qquad (2\text{-}3)$$

Fig. 2-4 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 with the other peaks. That is shown in Fig. 2-4 (c). The reasons are that the first formant, F$_1$, is damped-oscillation and the glottal pulse is asymmetric for the zero level. So 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 G-peak as the first peak and do side-peak as remainings.

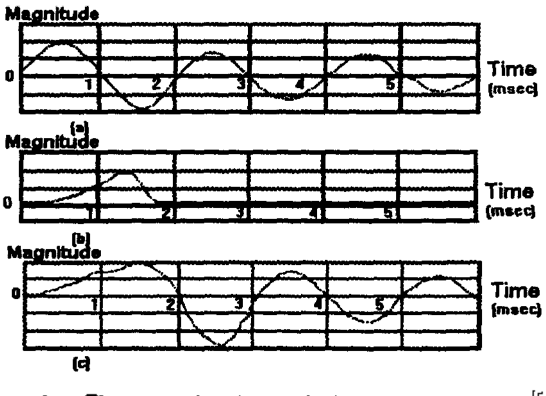


Fig. 2-4. The approximation analysis for voiced speech [5
(a) h(n) : impulse response of the approximated vocal tract
(b) g(n) : glottal waveform
(c) s(n) : voiced speech waveform by h(n) * g(n)

Ⅲ. Detection of the GCI by G-peak

The G-peak is the first peak of signal that convolved glottal waveform with amplitude of formants. 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 higher than that of side-peaks.

Because speech signal is convolved with many formants and glottal-pulse, it is very difficult to detect only G-peak in the voiced speech waveforms. Also, formants and G-peaks of speech, it is desirable to remove the higher formants of speech signal. To do that, the voiced speech is passed through the low-pass filter as following equation.

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

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 cutoff frequency of LPF, f_T, must be varied in each frame. In this paper, resultly, we take cutoff frequency of the filter by using the properties of G-peak. Because the ZCI of G-peak in a pitch is longest, the detected maximum ZCI becomes that of G-peak. Before finding the maximum ZCI, we must take the zero-crossing point, $Z_c(i)$, of the signals. Then, ZCI(i) is to subtract $Z_c(i)$ from $Z_c(i+1)$ as follows :

$$ZCI(i) = Z_c(i+1) - Z_c(i). \qquad (3\text{-}2)$$

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

$$N \doteqdot \max\{Z_c(0),\ Z_c(1),\ \cdots,\ Z_c(M-1)\}. \qquad (3\text{-}3)$$

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

The signal passed through the filter, S'(i), is shown in Fig. 3-1(b). In this figure, the G-peak in a pitch may be distinguished properly to the side peaks. Since S'(i) is asymetrical for ground, to remove the side-peaks, the threshold level for the G-peak can be taken by the maximum of side-peaks.

Hence, we decide the threshold value, L_{Th}, for extracting the G-peak in the frame as follows :

$$L_{Th} = |\min(S'(i))|, \qquad i = 0, 1, \cdots, L-1. \qquad (3\text{-}4)$$

Where L is the number of samples in one frame.

The peaks over the threshold are the G-peaks in the frame. Fig. 3-1(c) shows the detected G-peak. Because GCI is defined as closure instant of the glottis, the end point of G-peaks become the GCI points as shown in Fig. 3-1(d).

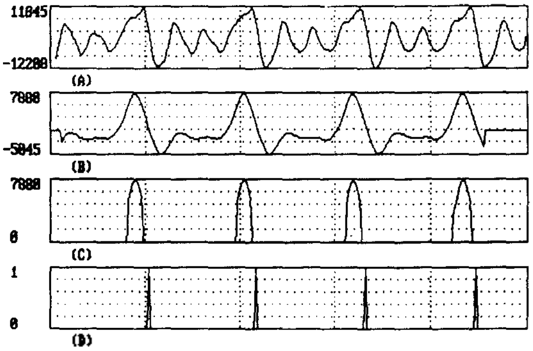


Fig. 3-1. The processing of GCI detection by property of the G-peak
(a) speech signals
(b) waveform through the variable LPF
(c) detected G-peak (d) detected GCI

Ⅳ. Experiment and Results

For the simulation, we use the IBM-PC/486DX2 (50) interfaced with A/D converter for input/output of speech signals. The sampling frequency is 8KHz and quantization level is 12bit/samples. The speech data composed of 5 Korean speakers' utterances(3 males and 2 females) and the following sentences were spoken each 5 times.

Utterance 1 :
/INSUNE KOMAGA CHUNJAE SONYUNWL
JOAHANDA/
Utterance 2 :
/JESUNIMKESEO CHUNJICHANGJOWI
KIOHUNWL MALSUMHASEOSSDA/
Utterance 3 :
/SOONGSILDAE JUNGBOTONGSINKONGHAKKWA
UMSEUNG SINHOCHURI YUNGU/
Utterance 4 :
/KAMSAHAMNIDA/

where the meaning of utterance 1 is "Insoo's young boy likes a genius kid", utterance 2 is "Jesus spoke of the lessons of the creation of the heavens and the earth", utterance 3 is "Speech signal processing team at the department of information and telecommunication, Soongsil University", and utterance 4 is "Thank you", spoken in Korea. The frame length for analysis is 256 samples provided conveniently to the simulation. The successive frames are overlapped with 128 samples.

In this paper, the speech waveforms are passed through the LPF that has the bandwidth of the filter as the maximum ZCI in a frame. We extract the G-peak over the threshold level that is the maximum of side-peak magnitude, and then detect the GCI from the G-peak.

We compared GCI detected in this method with eye-checked GCI as standard, and used the environment as the SNR of clean, 20dB, and 0dB. We took into account the range within 1ms between eye-checked and detected GCI. Table. 4-1. illustrates the performance scores for each utterance.

Table. 4-1. Performance score of GCI by eye-check

Utterance	Score(%)		
	Clean	20dB	0dB
1	97.2	96.1	94.7
2	98.1	96.4	95.0
3	97.5	95.9	94.1
4	98.9	97.5	95.3
Average	97.9	96.5	94.8

Ⅴ. Conclusion

It is important to exactly detect the GCI(Glottal Closure Instant) in the speech signal processing. The detection of GCI is applied to analyze the speech waveform to the pitch synchronously, to recognize with removing the personality of speakers, and to synthesize with high quality. It is difficult to detect the GCI for wide range of speakers. The problem is due to many significant residual pulses occuring around the GCI from both the input pulse and the large prediction error.

In this paper, we proposed a new algorithm for GCI detection by using the G-peak. The speech waveforms are passed through the LPF of variable bandwidth. Then, the GCI's of voiced speech are detected by the G-peak based on the filtered signals. We took into account the range within 1 ms between eye-checked and detected GCI. We have achieved the results as 97.9% in the clean speech, 96.5% in 20dB SNR, and 94.8% in 0dB SNR, respectively.

We obtained the fact that the proposed method has the high performance score in noise environment, and detected accurately the GCI for wide range of speakers and for all vowel signals.

Reference

1. L.R.Rabiner and R.W.Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
2. J.D.Markel and A.H.Gray, jr., *Linear Prediction of Speech Signals*, Springer-Verlag, 1976.
3. Dale E.Veenemean and Spencer L.Bement, "Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-33, No.2, pp.369-377, April 1985.
4. Yan Ming Cheng and Douglas O'Shaughnessy, "Automatic and Reliable Estimation of Glottal Closure Instant and Period," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.37, No.12, December 1989.
5. Hans Werner Strube, "Determination of the instant of glottal closure from the speech wave," J.Acoust. Soc.Am., Vol.56, No.5, pp.1625-1629, November 1974.
6. A.E.Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," J.Acoust.Soc.Am., Vol.49, pp.583-590, 1971.
7. L.R.Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," IEEE Trans on Acoustics, Speech, and Signal Proc., Vol.ASSP-26, No.1, pp. 24-33, February 1977
8. Myungjin BAE, Jaeyeol LIM, and Souguil ANN, "The Study of Extraction of the Energy of Speech Signal Using the G-peak in Speech Processing Model," Journal of the Korean Institute of Telematics and Electronics, Vol.24, No.3, pp.381-386, 1987.
9. D.Y.Wong, J.D.Markel, and A.H.Gray, Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust, Speech, Signal Processing, Vol.ASSP-27, pp.350-355, Aug. 1979.
10. Myungjin BAE and Souguil ANN, "A study on the fundamental frequency extraction of speech signals using second order rundown method," Seoul National University, MA Paper, Jan, 1983.
11. G.Fant, *Acoustic Theory of Speech Production*, Gravenhange, The Netherlands : Mouton, 1960.
12. Haegoon LEE, Myungjin BAE, and Uncheon LIM, "The Pitch Beginning Point Extraction Using G-peak from the Speech Production Model," Journal of the '93 Korean Signal Processing Conference, Vol.6, No.1, pp.58-61, 1993.

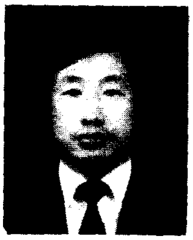
▲Hong Keum

Hong Keum was born on February, 5, 1970.
She received the B. S. degree in Electronic Engineering from Soongsil University, Seoul, in 1993. She is currently enrolled in a M.S. degree of Soorgsil University. Her research interests in clucle speech analysis, speech coding.

▲Daesik Kim

He received the B.S. degree in electronics engineering from Ho Seo University, and the M.S. degree in electronics engineering from soongsil University, in 1987 and 1989, respectrely. He is currently workrg at Lab. of Speech Communication of SoongSil University.

▲Young-Il Kim



Date of borm ; Feb, 25, 1954

1984 : Dept. of Electronic Engineering, Seoul Industrial Univ. (B.E.)

1988 : Dept. of Computer Science, Chung-ju Univ. (M.E.)

1993 : Lecturer of the Dept. of Electronic Engineering, Oson College

Interesting Research Area : Signal processing and Acoustic System Design

▲Myung Jin Bae

Vol.11, No.1E