On a Cepstral Pitch Alteration Technique for Prosody Control in the Speech Synthesis System with High Quality

*Kyu-Hong Kim, **Seong-Joon Baek, and ***Myung-Jin Bae

Abstract

In the area of the speech synthesis techniques, the waveform coding methods maintain the intelligibility and naturalness of synthetic speech. In order to apply the waveform coding techniques to synthesis by rule, we must be able to alter the pitches of synthetic speech. In this paper, we propose a new pitch altering method that compensates phase distortion of the cepstral pitch alteration method with time scaling method in the time domain. This method can remove some spectrum distortion which is occurred in conjunction point between the waveforms. For performance test the spectrum distortion rate was used as objective criterion and the MOS(Mean Opinion Score) was used as subjective criterion. As a result, the spectrum distortion and MOS are obtained by 0.66% and 3.9, respectively.

I. Introduction

Recently, owing to the rapid progress of VLSI technology, the 64 Mbit memory size per chip package is available in market. For the 32 kbps ADPCM waveform coding, such a long speech data as a half hour lasting speech can be stored by using one 64 Mbit chip. This makes the improvement of speech quality more important target than the reducing of memory size. The waveform coding method or the hybrid coding method, also, is preferable to the speech synthesis techniques for high quality. Although, for a long time, the waveform coding method and the hybrid coding method have been used for sentence based synthesis in synthesis technique by analysis, they are not proper to syllable or phoneme based synthesis techniques, because of the difficulty in controlling the excitation source. Even when they are used for word or demi-syllable based synthesis, different data are used even for same word according to the word connected to it. However, if we can alter the pitch period on speech waveform, the waveform coding techniques for the synthesis by rule is relatively good method to maintain the naturalness and the intelligibility comparable to the original speech.

According to processing domain, pitch alteration method is classified into three domains; time domain, frequency domain and time-frequency hybrid domain. There are multi-pulse method and pitch halving method in time domain. To alter the pitch period, Caspers and Atal proposed the method in which zeros are inserted or the data is deleted between pulses on MPLPC[5]. However, because the pulse train on MPLPC is related to pitch and formant, serious spectrum distortion occurs. Varga and Fallside had proposed the pitch extension method by LPC coefficients, which also causes serious spectrum distortion because they simply deleted a part of waveform when shortening the pitch[6].

Since formant variation causes an effect on the characteristics of vocal tract filter, some message information is lost and if the phase information is not kept, spectrum distortion on phoneme occurs because of the large variation in level. Generally, while the pitch altering on time domain causes large spectrum distortion and less phase distortion, the pitch altering on frequency domain causes less spectral distortion and large phase distortion. Therefore, we can get spectrum amplitude by altering the pitch on the frequency domain and then compensate the phase distortion of that on the time domain, when we want to minimize the distortion which comes from the pitch alteration.

In this paper, we propose a new pitch altering method in which pitch-altered waveform can be obtained by combining the pitch data which come from the cepstrum analysis and the phase data which come from the time scaling pitch control method.

II. Cepstral pitch alteration method

Unlike in the source coding method, the pitch variation of the speaker must be known prior to change the pitch period in the waveform speech coding. This comes from the fact that the variations of the accent and

* Information and Communication University.
** Applied Electronics Lab, S.N.U
** Soongsil University.

Manuscript Received : February 22, 1999

the emotion of a speaker result in the variation of the pitch period around the average value of that. Especially, since the waveform coding method conserves the characteristics of a speaker and the message informations, its intelligibility is relatively good. So, it is needed to alter the pitch period according to the average pitch period which mainly appears in the speech signal of the speaker. Therefore, the precise pitch detection must be carried out prior to changing the pitch.

From the result of cepstral analysis for the voiced speech, the combined contributions of vocal tract, glottal pulse and radiation appear on the lower part of quefrency domain and decay rapidly for large quefrency. The remarkable peak corresponding to the excitation source appears around the pitch period on the higher quefrency domain. So, by inserting lifter around the pitch where the cepstrum decay to zero on the quefrency domain, we can separate the formant components and the fundamental informations. This is called as the cepstral analysis method[1].

Speech signal can be separated into magnitude component and phase component by Fourier transform. So, the magnitude component of the Fourier transformed speech signal is as follows:

$$S(k) = \int_{-\infty}^{\infty} s(n)\ e^{-j\frac{n}{2\pi N}k} dn \tag{1}$$

$$M(k) = 10\log S^2(k) \tag{2}$$

To control the pitch in frequency domain, spectrum scaling is used. Spectrum must scale on the speech excitation spectrum. Thereby, the separation of component is performed before pitch alteration by the cepstral analysis.

If the formant components, S*(k), extracted by cepstral analysis are subtracted from M(k) as Equation (3), the flattened harmonics spectrum could be separated:

$$S_P(k) = M(k) - S^*(k) \tag{3}$$

Where Sp(k) is the flattened harmonics spectrum. For this signal, the scaling rate in frequency domain is the inversion of the scaling coefficient of time axis.

$$\widehat{S_P}(k) = S_P(k \times \rho^{-1}) \tag{4}$$

$$(k = 0, 1, 2, 3, \ldots, N-1)$$

In Equation (4), ρ^{-1} represents the frequency scaling rate, and $\widehat{S_P}(k)$ expresses the changed harmonics spectrum. It must decrease the interval of the fundamental frequency by ρ^{-1} for expanding pitch, and increase by ρ^{-1} for compressing pitch.

Since the effect depending on the kind of window is serious, the beginning point of window has to be synchronized to the exciting point of the glottal pulse. For this, the phase information of the waveform must be kept unchanged while changing the pitch period, so, time domain pitch extraction is desirable. In this paper, we adopt the area comparison method[4] in time domain. However, since the automatic pitch extraction is not positively necessary when editing the waveform for synthesis, semi-automatic pitch extraction and manual pitch extraction also may be a good adoption[7][8].

III. Phase compensation

In the cepstrum pitch alteration method proposed previously in [7], how phase information can be kept unchanged is the unresolved problem. So, we propose the phase compensation method in which we use the time domain pitch altering method with the cepstrum pitch alteration method at the same time.

Prior to control the pitch in time domain, voiced speech signal is passed through the low pass filter(LPF) represented as a following Equation (5) with a cut-off bandwidth as a pitch period.

$$s'(n - \frac{N}{2}) = \sum_{i=0}^{N-1} s(n-1) \tag{5}$$

Where N is the cut-off bandwidth interval of LPF, because the cut off frequency, fT, equals fS/N. For the harmonics above the fundamental frequency is removed from the signal, the LPFed signals are similar to excitation source of the voiced signals. Now, the signal is scaled at time axis as follows:

$$\widehat{s(n)} = s'(n \times \rho) \tag{6}$$

Where $\widehat{s(n)}$ is the scaled signal in time domain. s'(n) is the low pass filtered signal. The scaling factor is ρ as follows:

$$\rho = \frac{P'}{P} \tag{7}$$

where P is a speaker's pitch and P' is an expected pitch. If ρ is smaller than 1, we would obtain the signal with compressed pitch, Reversely if ρ is larger than 1, we would obtain the expended pitch. Then, the FFT is applied to the signal scaled at time axis.

As represented so far, the phase information is obtained from the FFT spectrum after we alter the pitch period of the speech in time domain by time scaling method, and then it is combined with the magnitude of the spectrum which is obtained by cepstrum pitch alteration method.

IV. Experimental results

The proposed algorithm has been implemented on the IBM PC(Pentium 233Mhz) with the 16-bit AD-DA converter. The speech signal is low-pass filtered at 4 kHz and sampled at 11 kHz. Five phoneme balanced Korean sentences are used as test data. Each sentence is pronounced 5 times by three males and two female speakers. The following sentences are used in our experiment:

Data 1. /INSUNE KOMAGA CHUNJAE SONYUNWL JOAHANDA/
Data 2. /JESUNIMKESEO CHUNJICHANGJOWI KIOHUNWL MALSUMHASEOSSDA/
Data 3. /SOONGSILDAE JUNGBOTONG SHIN-KWA UMSENG SINHOCHURI YUNGUTEEMIDA/
Data 4. /KAMSAHAMNIDA/
Data 5. /May I Help You/

One analyzed frame consists of 512 samples. First, the beginning point of pitch period is obtained by using the area comparison method to get one pitch interval which is needed for synthesis by rule in waveform coding. After repeating this interval to get a frame which consists of 512 samples, we altered the pitch period. Figure 1 is the block diagram proposed in this paper.

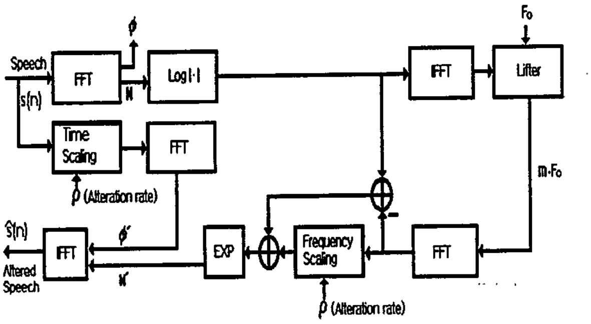


Figure 1. A block diagram of proposed pitch alteration technique.

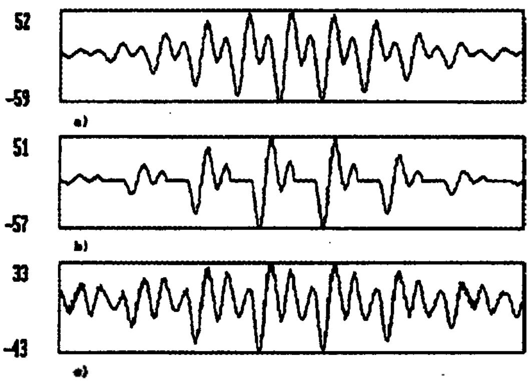


Figure 2. An example of pitch period extended by 50% using the cepstrum pitch alteration method.

The examples depicted in Figure 2 are the results of Data 1, when the pitch period is altered by 150%. Figure 2(a) shows the original speech signal and Figure 2(b) shows the resultant signal after the pitch is scaled by using the zero inserting and halving method in time domain. The result in Figure 2(c) is obtained by using the proposed method without losing the phase information.

As shown in Figure 2(b) and Figure 2(c), the phase information of both are same. So, we can obtain the pitch altered signal whose beginning point of one pitch interval is synchronized to that of the original signal. From the fact noted above, we can minimize the phase distortion which is generated around the conjunction point of adjacent waveforms in synthesis by rule by using waveform coding.

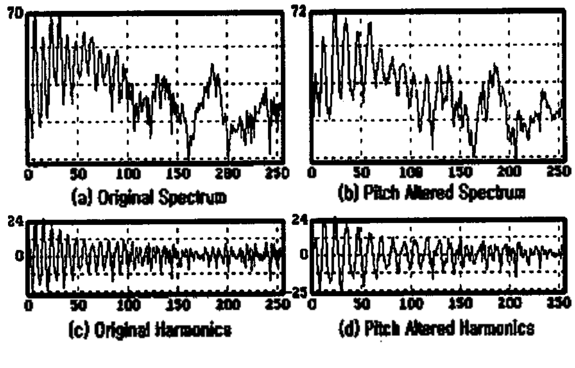
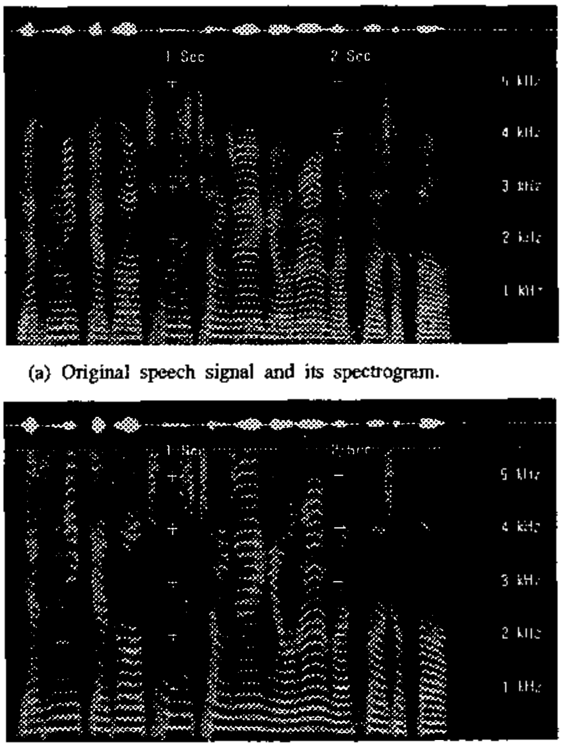


Figure 3. An example of pitch compressed by 70% by using the cepstrum pitch alteration method.



(a) Original speech signal and its spectrogram.

(b) Pitch altered speech signal and its spectrogram

Figure 4. The spectrograms of original speech signal and pitch altered speech signal by 70%.

Figure 3 shows an example of pitch compressed by 70% by using the cepstrum pitch alteration method. Figure 3(a) shows original spectrum, Figure 3(b) shows pitch altered spectrum by using the cepstrum pitch alteration method, Figure 3(c) shows original harmonics and Figure 3(d) shows pitch altered harmonics. As shown in Figure 3(a) and Figure 3(b), the formant envelope can be maintained when the pitch is altered.

The spectrograms depicted in Figure 4 are the results of computer simulation for Data 2. Figure 4(a) shows original speech signal and its spectrogram and Figure 4(b) shows pitch altered speech signal and its spectrogram. For performance test the spectrum distortion rate was used as objective criterion and the MOS(Mean Opinion Score) was used as subjective criterion.

Table 1. Spectrum distortion rates in the cepstral pitch alteration method.

Alteration Rate	Female (%)	Male (%)	Average (%)
90% → 111%	0.23	0.19	0.21
80% → 125%	0.42	0.35	0.39
70% → 142%	0.73	0.53	0.63
60% → 166%	1.10	0.71	0.91
50% → 200%	1.54	0.82	1.18
Average	0.80	0.52	0.66

Table 2. The result of MOS test in the cepstral pitch alteration method.

Alteration Rate	MOS	
	Conventional method	Proposed method
90% → 111%	4.0	4.2
80% → 125%	3.8	4.1
70% → 142%	3.5	4.0
60% → 166%	3.1	3.8
50% → 200%	2.8	3.2
Average	3.4	3.9

Table 1 shows the spectrum distortion rate of the resultant spectrum obtained by using the cepstrum pitch alteration method, which is compared with the spectrum of original speech. In this table, we first compress the pitch of the speech signal by fixed percentage and then expand the signal to compare it with the original speech, that is to say, lengthen the pitch on the frequency domain. From the table, we can find that the spectrum distortion of female sound, i.e. high frequency sound, is relatively high. However, the overall values on the table are much low. Table 2 shows the result of MOS test. From the MOS test, the average MOS of conventional method[7] and proposed method was 3.4 and 3.9 respectively.

V. Conclusions

Speech synthesis techniques are classified into three groups; waveform coding, source coding and hybrid coding. Waveform coding and hybrid coding methods are mainly used to synthesis method by analysis for a long time so far, because of the difficulty of pitch altering makes them improper to synthesis by rule. However, if it is possible to alter the pitch period when the waveform coding is used, synthesis by rule is available with maintaining good intelligibility and naturalness comparable to the original speech.

In this paper, we proposed the new pitch altering method, in which the magnitude spectrum of pitch altered speech signal is obtained by using the cepstral pitch altering method and the phase compensation is performed on that by using the time scaling method. Since we alter the pitch period over the magnitude spectrum flattened on the frequency domain where the formant informations almost does not exist, we can minimize the magnitude spectrum distortion. And we can minimize the phase spectrum distortion which is generated in conjunction point of two analyzed frame when using synthesis by rule in waveform coding.

As a result, the spectrum distortion was 0.66% and the MOS was 3.9.

References

1. L.R. Rabiner & R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, New Jersey, 1978.
2. M.S. LEE, M.J. BAE, J.H. LEE and S.G. ANN, "On Realizing the Predictor for the Waveform Coding of Speech Signals by using the Dual First Order Autocorrelation," *J., Acoust., Soc., Korea,* Vol.11, No.1E, pp.23-29, JULY 1992.
3. M.R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," *IEEE, Trans., Acoust. Speech, Signal Processing,* Vol.29, No.3, pp.374-390, June 1981.
4. M.J. BAE and S.G. ANN, "the High Speed Pitch Extraction of Speech Signals Using The Area Comparison Method.," KITE, Vol.2, No.2, pp.101-105, Feb., 1985.
5. B.E. Caspers and B.S. Atal, "Changing Pitch and Duration in LPC Synthesised Speech using Multipulse Excitation," *J. Acoust. Soc. Amer.,* Vol.73, No.1, pp.55, Spring, 1983.
6. A. varga and F. Fallside, "A Technique for Using Multipulse Linear Predictive Speech Synthesis in Text-to-speech Type System," *IEEE signal processing,* Vol.ASSP-35, No.4, pp.586-587, APRIL 1987.
7. M.J. BAE and M.S. LEE, "On a Pitch Change of the Waveform Coding by the Cepstrum Analysis of Speech Waveforms," *J., Acoust., Soc., Korea,* Vol.11, No.4, pp.14-21, August 1992.
8. M.J. BAE, H.S. YOON and S.G. ANN, "On Altering the Pitch of Speech Signals in Waveform Coding -Alteration Method by the LPC and Pitch Halving-," *J., Acoust., Soc., Korea,* Vol.10, No.5, PP.11-19, Oct. 1991.
9. M.J. BAE and S.H. LEE, "On a Cepstral Technique for Pitch Control in the High Quality Text-To-Speech Type System," 39'th Midwest Symposium on Circuits and Systems, Procedding of MWSCAS'96, August 18-21, 1996.

▲ Kyu-Hong Kim

Kyu Hong Kim was born in Inchon, Korea, on september 3, 1974. He received the B.S. and M.S. degree in Information and Telecommunication Engineering from Soongsil University, Seoul, Korea, in 1997 and 1999, respectively. From 1999 to present, he is a Ph.D. candidate in Multimedia Information and Communication Engineering at ICU(Information and Communications University) in Taejon, Korea. His research interests include speech recognition, speech coding, speech synthesis, and signal processing.

▲ SeongJoon Baek

Seong Joon Baek was born on Jan. 16 in 1967. He received the B.S. degree, the M.S. degree, and Ph.D. degree in electronics engineering from Seoul National University. Currently, he is with the applied electronics lab. in Seoul National Univ. His resrarch interests include speech coding, speech/speaker recognition and signal processing. He is a mrmber of Acoustical Society of Korea.

▲ Myung-Jin Bae

Myung Jin Bae was born in Kyungsangbukdo, Korea on May, 20, 1957. He received the B.S. degree in electronics engineering from Soongsil University and the M.S. degree in electronics engineering from Seoul National University in 1981 and 1983, respectively. From the same university, he received the Ph.D. degree in electronics engineering, too, in 1987. He has been professor for department of information and telecommunication of Soongsil university, located at Seoul in Korea, ever since 1992. His research interests include speech coding, synthesis, and speaker recognition.