On a Pitch Alteration Technique in Excited Cepstral Spectrum for High Quality TTS

JongDeuk Kim, SeongJoon Baek, and MyungJin Bae*

Dept. of Telecommunication Engineering, Soongsil University
1-1, Sangdo-5dong, Dongjak-gu, Seoul 156-743, KOREA
jdkim@assp.soongsil.ac.kr, mjbae@saint.soongsil.ac.kr

*School of Electrical Engineering, Seoul National University,
Seoul 151-742, KOREA

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 or hybrid 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 alteration method that minimizes the spectrum distortion by using the behavior of cepstrum. This method splits the spectrum of speech signal into excitation spectrum and formant spectrum and transforms the excitation spectrum into cepstrum domain. The pitch of excitation cepstrum is altered by zero insertion or zero deletion and the pitch altered spectrum is reconstructed in spectrum domain. As a result of performance test, the average spectrum distortion was below 2.29% while that of conventional method is 2.47%.

1. INTRODUCTION

Speech synthesis has been studied for several decades. Speech synthesis methods may have two conflict goals: compression rate and speech quality. 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 the synthesis by rule, 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.

For time domain methods, Multi-Pulse method and Pitch Halving method were proposed. Caspers and Atal proposed a method that inserts or eliminates zero value between the multi pulses[3]. But, because the position of pulse is determined optimally, changing the position of pulse derives serious spectrum distortion. Also, Varga and Fallside proposed a pitch extension method using LPC coefficients[4]. This method also causes spectrum distortion because it simply deletes a part of waveform in pitch compression. The pitch halving method extends the waveform to double of expected pitch period by LPC synthesis and halves the period of extended waveform by decimation[5]. Because this method is performed only in time domain, the intelligibility is lessened by the spectrum distortion.

For frequency domain methods, there is the pitch alteration method that separates the spectrum of speech signal into formant component and excitation component and alter the pitch by scaling the excitation component [6]. The major drawback is the preservation of phase in time domain.

For time-frequency (hybrid) domain methods, Bae and Lee proposed a pitch alteration that inserts or eliminates zeros where the value of cepstrum is around zero[7]. This method cannot preserve a phase. Takagi and Miyasaka suggested a method compensating spectrum distortion using an LPC envelope in spectrum domain[8]. Such a compensating method

cannot handle all voiced sounds because it uses mainly peaks of an LPC spectrum envelope.

In this paper, we propose a new pitch alteration method that changes the pitch in cepstrum domain by zero insertion or zero deletion after splitting the spectrum into excitation and formants in spectrum domain for the minimization of spectrum distortion.

2. PROPERTY OF CEPSTRUM

Vocal track, glottal wave and radiation informaion are shown in the lower part of cepstrum and excitation information is shown in the higher. The following Lifter window function, *l(n)*, that is called frequency-invariance linear filter selects the necessary component from cepstrum.

$$\begin{aligned} l(n) &= 1, \quad |n| < n_0 \\ &= 0, \quad |n| \geq n_0 \end{aligned} \tag{2-1}$$

Where, n_0, that is less than the period of fundamental frequency, N_p, is chosen.

Flattened log function can be obtained by applying the window function to cepstrum. This flattened log spectrum represents resonance of input speech frame and accords with formant frequencies fundamentally. Also, The lower part of cepstrum that represents the properties that are mixed with vocal cord, glottal pulses and radiation components respectively is reduced rapidly as quefrency increases.

On the other hand, if the lifter window function, *l*(n), of cepstrum is selected as following so as to make excitation selected, the higher part of cepstrum is emphasized. It is thereby possible to estimate fundamental frequency and voiced speech.

$$\begin{aligned} l(n) &= 0, \quad |n| < n_0 \\ &= 1, \quad |n| \geq n_0 \end{aligned} \tag{2-2}$$

In case of voiced speech the peaks on cepstrum are shown at the position of fundamental period of speech frame but in case of unvoiced they are not. It is possible to make a decision whether the present speech frame is voiced or unvoiced speech using those properties and also possible to detect a fundamental period of voiced speech.

Spectrum distortion needs to be minimized while keeping phase in time domain. Formant spectrum distortion results in loss of meaning because it changes filter information. On the other hand, phase distortion causes loss of naturalness due to amplitude fluctuation between adjacent frames.

We propose a new pitch alteration method that attempts to minimize spectrum distortion.

3. PITCH ALTERATION BY CEPSTRUM ANALYSIS OF FLATTENED EXCITATION SPECTRUM

The pitch alteration methods by component separation split the speech spectrum into excitation and formants and alter the excitation by scaling. However the insertion or deletion of harmonics in higher band leads the spectrum distortion as Fig. 3-1. This spectrum distortion causes the deterioration of synthetic speech.

For minimizing this spectrum distortion we alter the pitch in cepstrum domain. The speech signal is transformed into magnitude spectrum and phase spectrum by Fourier transform. A definition of the Fourier transform is

$$S(K) = \int_{-\infty}^{\infty} s(n) e^{-j\frac{n}{2\pi N}k} dn \tag{3-1}$$

The log magnitude spectrum and the phase spectrum is computed as Eq. 3-2 and Eq. 3-3.

$$M(K) = 10 \log S^2(K) \tag{3-2}$$

$$\phi(K) = \tan^{-1} \frac{\mathrm{Im}[S(K)]}{\mathrm{Re}[S(K)]} \tag{3-3}$$

Re[S(K)] and Im[S(K)] represent the real part and the imaginary part of spectrum respectively.

For splitting the magnitude spectrum into excitation and formants, the approximated formants spectrum, H(K), is acquired by applying a Lifter function as Eq. 3-4.

$$H(K) = \frac{1}{K_0} \sum_{i=-\frac{K_0}{2}}^{\frac{K_0}{2}} M(K-i) \tag{3-4}$$

K_0 represents a fundamental frequency. The flattened excitation spectrum, E(K), is computed as follow.

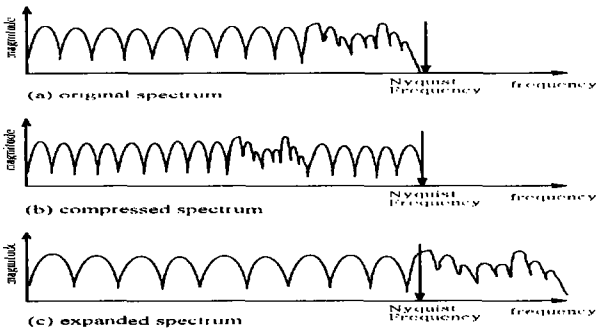

Figure.3-1. The spectrum distortion by Scaling in pitch Alteration

$$E(K) = M(K) - H(K) \qquad (3\text{-}5)$$

This excitation spectrum is transformed into cepstrum by Inverse Fourier transform. The cepstrum values of excitation are nearly zero except pitch pulse, because the cespstrm includes only excitation component. For changing the pitch adequate number of zeros are inserted or deleted between zero quefrency and pitch pulse. The inserting or deleting of zeros can little affect the synthetic speech. The excitation cepstrum with altered pitch is transformed into spectrum again by Fourier transform. The log magnitude spectrum with altered pitch is sum of approximated formants spectrum and excitation spectrum with altered pitch as follow.

$$M'(K) = H(K) + E'(K) \qquad (3\text{-}6)$$

Now, M'(K) is passed through an exponential function to make a magnitude spectrum. Then magnitude spectrum and pitch altered phase spectrum are fed into Inverse Fourier transform for the final speech waveform with altered pitch.

4. PITCH DETECTION

We must know the usual pitch of speaker before pitch alteration since the pitch variation due to intonation or emotion is relative to the usual pitch of speaker. So, when we synthesize the speech with emotion or intonation, we must alter the pitch relative to the usual pitch. Thus, we find an exact pitch period before pitch alteration.

Many methods for pitch detection have proposed until now. It can be classified into three category; time domain, frequency domain and time-frequency (hybrid) domain method. We find the pitch by the area comparison method(ACM) in time domain[9]. It needs not automatic detection in editing the waveform for synthesis. It may be semi-automatic or manual method.

5. EXPERIMENTAL RESULT

Our method was implemented on an IBM PC/pentium (150MHz) with a 16-bit A/D converter. The sampling frequency was 11kHz. The speech data was sampled from 4 Korean speakers, two males and two females ones. The sentences used in our experiments are as followings.

Utterance1) /IN SU NE KOMANUN CHUNJE SONYUNWL JOAHANDA/
Utterance2) /JESUNIMKESEO CHUNJI CHANGJOWI KYOHUNWL MALSUMHASEOSSDA/
Utterance3) /YEOGINUN UMSEONGHAPSEONG YUNGUSIL IMNIDA/
Utterance4) /KONG IL I SAM SA O RUK CHIL PAL GU /
Utterance5) /KAM SA HAM NI DA/

One analyzed frame consists of 256 samples. We decide the pitch of the speech signal by using the area comparison method. Coincidentally, we apply the hamming window to speech signal and transform into frequency domain by FFT.

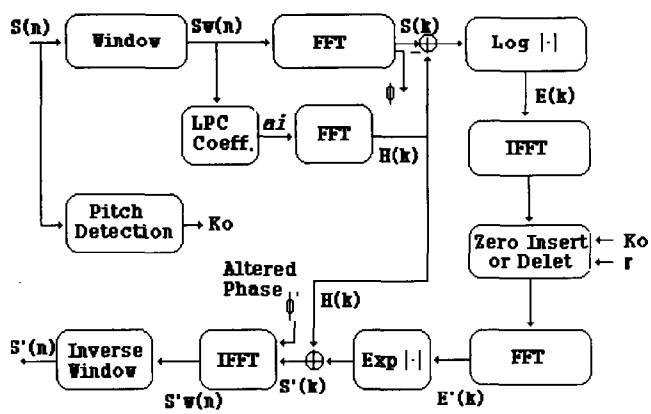

Figure. 5-1. The block diagram of proposed method.

To obtain the approximated formant spectrum we perform the log operation and liftering with cutoff frequency of fundamental frequency and, then, subtract the approximated formant spectrum from log spectrum for obtaining excitation spectrum.

The excitation spectrum is transformed into cepstrum by IFFT and altered pitch by zero inserting or zero deleting

between zero quefrency and pitch pulse. The excitation cepstrum with altered pitch is transformed into spectrum by FFT and the pitch altered spectrum is reconstructed by adding approximated formant spectrum and pitch altered excitation spectrum.

	Conventional method			Proposed method		
	Male	Female	Aver	Male	Female	Aver
120%	1.67	2.03	1.85	1.49	1.83	1.66
140%	2.04	2.32	2.18	1.75	2.15	1.95
160%	2.18	2.50	2.34	1.84	2.38	2.11
180%	2.67	2.98	2.83	2.52	2.74	2.63
200%	2.84	3.43	3.14	2.82	3.38	3.10
Aver	2.28	2.65	2.47	2.08	2.50	2.29

Table 5-1. The comparison of spectrum distortion.

Finally, the speech signal with altered pitch is reconstructed by the exponential function and IFFT. Fig.5-1 is the block diagram of proposed method. For performance test, we measured the spectrum distortion rate of synthetic speech with altering the pitch by 120%, 140%, 160%, 180%, 200%.

Table 5-1 shows the spectrum distortion rate and Fig.5-2 is the result of 120% pitch alteration. As a result of performance test, the average spectrum distortion was below 2.29% while that of conventional method is 2.47%.

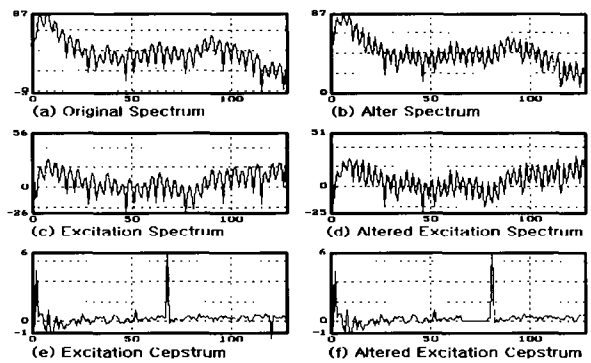

Figure.5-2. The result of 120% pitch alteration.

6. CONCLUSION

Speech coding is classified into three categories: waveform coding, source coding and hybrid coding. To obtain the synthetic speech with high quality, the synthesis using waveform coding is desired. However, it is difficult to apply for waveform coding to synthesis by syllable or phoneme unit, because it does not divide the speech into excitation and formant component. Thus, it is required to alter the excitation in waveform coding for applying waveform coding to the synthesis by rule.

In this paper, we propose a new pitch alteration method that minimizes the spectrum distortion by using the behavior of cepstrum. This method splits the spectrum of speech signal into excitation spectrum and formant spectrum and transforms the excitation spectrum into cepstrum domain.

The pitch of excitation cepstrum is altered by zero insertion or zero deletion and the pitch altered spectrum is reconstructed in spectrum domain.

As a result of performance test, the average spectrum distortion was below 2.29% while that of conventional method is 2.47%.

REFERENCE

[1] L.R. Rabiner and R.W. Schafer, *Digital Processing of speech Signals*, Prentice-Hall, 1978.

[2] A.M. Kondoz, *Digital Speech*, John Wiley & Sons, 1994.

[3] B.E. Caspers and B.S. Atal, "Changing Pitch and Duration in LPC Synthesized Speech using Multipulse Excitation," J. Acoust. Soc. Amer., Vol.73, No.1, pp.55, spring 1983.

[4] 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.

[5] M. BAE, H. YOON, S. 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.

[6] T.F. Quatieri, R.J. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech," IEEE Trans., Signal Processing, Vol.40, No.3, pp.497-510, March 1992.

[7] 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, Proceeding of MWSCAS'96, pp.803-806, August 18-21, 1996.

[8] T. Takagi and E. Miyasaka, "A speech prosody conversion system with a high quality speech analysis-synthesis method. "Proc. EUROSPEECH'93, pp.995-998. September 1993.

[9] M.J. Bae and S.G. Ann, "The high speed pitch extraction of speech signal using the area comparison method," KITE, Vol.2, No.2, pp.101-105, Feb. 1985.