고음질 운율조절용 시간-주파수 혼성영역 피치변경법

On a Pitch Alteration Technique in Time-Frequency Hybrid Domain for High Quality Prosody Control of Speech Signal

이 상 효*, 배 명 진*
(Sang-Hyo Lee*, Myung-Jin Bae*)

요 약

음성합성분야에서 파형부호화 합성방식은 합성음의 자연성과 명료성을 유지할 수 있다. 그렇지만 법칙에 의한 합성방식에 적용하려고 하면 운율조절을 위해 음성의 피치를 변경해야만 한다. 우리는 본 논문에서 시간영역에서 시간축조절 피치변경법에 의해 켑스트럼 피치변경법의 위상왜곡을 보상하는 시간-주파수 혼성형 피치변경법을 새로이 제안하였다. 이 방법은 연속 프레임에서 파형들간의 연결점에서 유발될 수 있는 위상스펙트럼 왜곡을 제거할 수 있고, 또한 200%의 피치변경에 대해서도 진폭스펙트럼의 왜곡이 1.18% 이하인 성능을 얻었다.

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, however, we must be able to alter the pitches for prosody control of synthetic speech. In this paper, we propose a new pitch alteration technique in time-frequency hybrid domain, that compensates phase distortion of the cepstral pitch alteration method with time scaling method in the time domain. This method can remove some phase spectrum distortion which is occurred in conjunction point between the waveforms in continued frames. Also, we can obtain little magnitude spectrum distortion below 1.18% for pitch alteration of 200%.

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[2], 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[7]. 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

*숭실대학교 정보통신공학과
접수일자: 1997년 3월 9일

when shortening the pitch[8].

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 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:

$$S(K)=\int_{-\infty}^{\infty} s(n)\ e^{-j\frac{\pi}{2\pi N}k}\,dn \tag{2.1}$$

$$M(K)=10\log S^2(K) \tag{2.2}$$

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 Eq. (2.3), the flattened harmonics spectrum could be separated:

$$S_P(K)=M(K)-S^*(K) \tag{2.3}$$

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

$$\hat{S}_P(K)=S_P(K\times\rho^{-1}) \tag{2.4}$$
$$(K=0,\ 1,\ 2,\ 3,\ \ldots,\ size-1)$$

In Eq. (2.4), ρ^{-1} represents the frequency scaling rate, and $\hat{S}_p(K)$ 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 in time domain[5]. 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[1][6].

Ⅲ. Phase Compensation

In the cepstrum pitch alteration method proposed previously in [9], 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 Eq. (3.1) with a cut-off bandwidth as a pitch period.

$$s'(n-\frac{N}{2}) = \sum_{i=0}^{N-1} s(n-1) \tag{3.1}$$

Where N is the cut-off bandwidth interval of LPF, because the cut off frequency, f_T, equals f_S/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:

$$s(\hat{n}) = s'(n \times \rho) \tag{3.2}$$

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

$$\rho = \frac{P'}{P} \tag{3.3}$$

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/586 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 SHINKWA UMSENG SINHOCHURI YUNGUTEEMIDA/
Data 4)/KAMSAHAMNIDA/
Data 5)/May I Help You/

Fig. 4-1 is the block diagram proposed in this paper. 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 according to processing in the block diagram.

Table 4-1 shows the 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.

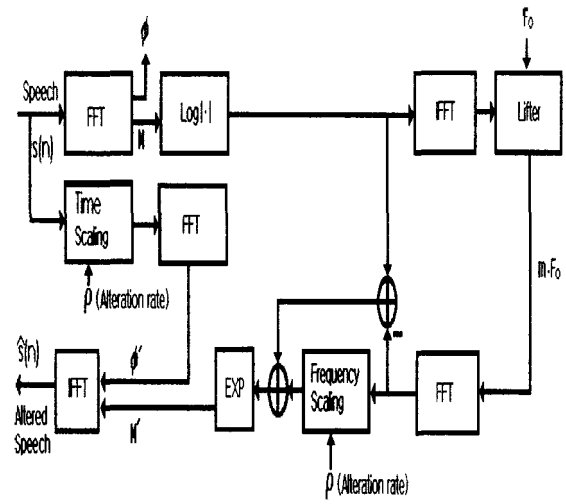

Figure 4-1. A block diagram of proposed pitch alteration technique.

Ⅴ. Conclusions

Speech synthesis coding 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 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 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 on using synthesis by rule in waveform coding.

Table 4-1. 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

Acknowledgement

The authors would like to thank the Korea Science & Engineering Foundation for their support of this work (Project No.:95-0100-22-01-3).

References

1. L.R. Rabiner & R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, 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.J. BAE, "The TTS Speech Synthesis Techniques", Proceedings of Korea Inst. Commun. Sciences, Vol. 11, No. 9, pp. 67-78, Sept. 1994.
4. M.R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," IEEE, Trans., Acoust. Speech, Signal Processing, Vol. ASSP-29, No. 3, pp. 374-390, June 1981.
5. 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.
6. H. HONG, G.R. BAEK, M.J. BAE, H.S. JANG, "Pitch Detection using Variable Bandwidth LPF", J., Acoust., Soc., Korea, Vol. 13, No. 5, pp. 77-82, October 1994.
7. 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.
8. 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.
9. 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.
10. 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.
11. M.J. BAE, "On a Pitch alteration Method using Scaling the Harmonics Compensated with the Phase for a Speech Synthesis," J., Acoust., Soc., Korea, Vol. 13, No. 6, pp. 91-97, December 1994.
12. M.J. BAE, W.C. LEE and S.B. IM, "On a Pitch Alteration Method by Time-axis Scaling Compensated with the Spectrum for High Quality Speech Synthesis," J., Acoust., Society, Korea, Vol. 14, No. 4, pp. 89-95, August 1995.
13. W.R. JO, M.J. BAE and D.S. KIM, "On a Pitch Alteration Technique in the V/UV Spectrum for High Quality Speech Synthesis Technique," J., Acoust., Society, Korea, Vol. 15, No. 6, pp. 99-103, December 1996.

▲이 상 효(Sang Hyo Lee) 1973년 5월 10일생

1996년 2월:숭실대학교 정보통신공학과 졸업(공학사)
1996년 3월~현재:숭실대학교 대학원 전기공학과 석사과정
※주관심분야:음성 신호처리, 음성 통신

▲배 명 진(Myung Jin Bae)
한국음향학회지 15권 6호 참조(1996. 12)