위상 보상된 고조파 스케일링에 의한 음성합성용 피치변경법

On a Pitch Alteration Method using Scaling the Harmonics Compensated with the Phase for Speech Synthesis

배 명 진*
(Myungjin BAE)

ABSTRACT

In speech processing, the waveform codings are concerned with simply preserving the waveform of signal through a redundancy reduction process. In the case of speech synthesis, the waveform codings with high quality are mainly used to the synthesis by analysis. Because the parameters of this coding are not classified as both excitation and vocal tract, it is difficult to apply the waveform coding to the synthesis by rule. Thus, in order to apply the waveform coding to synthesis by rule, it is necessary to alter the pitches.

In this paper, we proposed a new pitch alteration method that can change the pitch period in waveform coding by dividing the speech signals into the vocal tract and excitation parameters. This method is a time-frequency domain method preserving the phase component of the waveform in time domain and the magnitude component in frequency domain. Thus, it is possible that the waveform coding is carried out the synthesis by rule in speech processing. In case of using the algorithm, we can obtain spectrum distortion with 2.94%. That is, the spectrum distortion is decreased more 5.06% than that of the pitch alteration method in time domain.

요 약

신호처리에서, 파형부호화법은 음성신호의 잉여성분을 감소시킴으로써 파형을 유지하는 부호화 방법이다. 음성 합성의 경우, 고음질의 파형부호화법은 주로 분석에 의한 합성법에 이용된다. 그러나, 파형부호화법은 여기 파라미터와 성도 파라미터로 분리하지 않고 처리하기 때문에 규칙에 의한 합성에 적용되기 어렵다. 따라서 파형부호화법을 규칙에 의한 합성에 이용하기 위해서는 피치변경이 필요하다.

본 논문에서, 우리는 파형부호화법에서 음성신호를 성도 파라미터와 여기 파라미터로 분리함으로써 피치 주기를 바꿀 수 있는 새로운 피치변경법을 제안한다. 이 방법은 시-주파수 혼성영역 방법으로 시간영역에서 파형의 위상성분과 주파수영역에서 파형의 진폭성분을 보존한다. 따라서 파형부호화법은 음성처리에 있어 규칙에 의한 합성을 할 수 있다. 본 논문에서 제안한 알고리즘을 이용한 경우, 단지 2.94%의 스펙트럼 왜곡만이 일어났다. 즉, 스펙트럼 왜곡이 시간영역에서의 피치변경법보다 5.06% 이상 감소되었다.

* Dept. of Telecommunication Engineering, Soongsil University
접수일자 : 1994년 9월 20일

Ⅰ. Introduction

According to the data unit of speech synthesis, speech synthesis method can be classified into three types : sentence synthesis, syllable synthesis and phoneme synthesis. Also, according to the method of transmission or storage for speech signals, the speech coding algorithms can be classified into three types : waveform coding, source coding and hybrid coding[1-3]. For the transmission efficiency or saving storage, the waveform coding is mainly used in base band coding.

In the waveform coding methods, the redundancy in speech waveform is reduced before it is transmitted through the transmission channel or storage medium. PCM, ADM, and ADPCM belong to this type[14]. These coding methods are not used for the syllable or phoneme synthesis methods which change the source, because the vocal tract filter information representing the meaning of message and the excitation information reflecting the personality and feeling of a person, are not separated in two parts in the processing procedure.

The source codings are very closely based on the speech production model. They separate the excitation information and the filter information in speech signals before these coding methods are realized. The methods that belong to this category are LPC, PARCOR, and LSP[15]. The application of source coding debase the naturalness and intelligibility of synthesized speech signals by using the divided components.

The hybrid coding methods have the merit of memory efficiency in the source coding and the merit of the naturalness and intelligibility in waveform coding. These methods are classified into RELP, VELP, MPLP, and CELP[16]. The filter information in the hybrid methods is generally coded by the waveform coding methods. Therefore it is difficult to apply these methods to the synthesis algorithm with unit of phoneme or syllable.

Recently, we want high quality in various areas of speech. For this, it is appropriate to use the waveform coding methods. But, in this case, we have problems that memory capacity is large and the pitch alteration is hard. The current technique is enough to overcome the problems of memory capacity. The solution of the other problem must be the pitch alteration of excitation source to apply the waveform coding to the synthesis by rule.

Ⅱ. Existing pitch alteration methods

The waveform coding and the hybrid coding have been applied to the sentence synthesis methods which use the synthesis by analysis. But, in reality, these are not used as syllable or phoneme synthesis methods on account of the difficulty of changing the pitch. Occasionally, the waveform coding or hybrid coding is used in a semi-syllable or a word synthesis. But, even if they are the same words, different data are used according to the following word. Therefore it is necessary to alter the pitch for synthesis by rule with high quality by using the waveform coding methods.

According to processing domain, pitch alteration method is classified into three domains : the time domain, the frequency domain and the time-frequency hybrid domain. There are Multi-Pulse method and Pitch halving method in time domain. Caspers and Atal had proposed the method that inserts or eliminates zero between the pulses in speech signals. But, Becasuse the pulse sequence on MPLPC have correlation between the pitch and the formant, the spectrum distortion is very large[4]. Also Varga and Fallside had proposed the pitch extension method using LPC coefficient. But this has a lot of spectrum distortion for the method leave out and flatten a part of speech waveform[5].

The pitch halving method in time domain, makes the waveform with the double of expected

period, and then halves the period of waveform [7]. Also, this method is performed only in time domain, the intelligibility is lessened by the occurence of spectrum distortion.

As the pitch alteration in frequency, there is the pitch alteration method that obtains the waveform with altered pitch by using separating the formants and harmonics of fundamental frequency in speech signals, and then scaling linearly the fundamental frequency. The major drawback of this method is the preservation of phase in speech signals[13].

As the time-frequency hybrid domain, there is the pitch changing method uses the feature of cepstrum. This eliminates or inserts zero in the part with zero cepstrum. But this method also has the problem of phase preservation[6].

If the spectrum's formant is changed on the pitch alteration, the filter information of vocal tract would be changed, so the information of opinion wouldn't preserve. Thereby enlarged level variation makes the unnatural connection between the phonemes.

Thus, in this paper, we propose the new pitch alteration method in time-frequency domain that preserves the phase characteristic in time domain and minimizes the distortion of speech spectrum in frequency domain. In the following section, we will explain the proposed pitch alteration method. And, in section 5 and 6, we present the experimental results and give the conclusion, respectively.

Ⅲ. The phase control in time domain

For pitch alteration in waveform coding, we have to know the pitch variation of speaker in advance, because the change of speaker's emotion and intonation cause the relative variation by the referenced main pitch. Especially, the waveform coding have a good intelligibility since the methods maintain an individuality and message information of speaker. Therefore, in the case of pitch alteration, it is necessary to change pitch based on the main pitch of speaker. Thus, first of all, we must detect the pitch exactly.

In this paper, we apply the area comparison method to pitch detection in time domain. The pitch alteration in frequency domain causes the loss of phase information. That is, pitch alteration brings about the variation of harmonics position. Therefore it is necessary to compensate the phase information.

To control phase in time domain, the voiced 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 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.

$$\hat{s}(n) = s'(n \times \rho) \tag{3-2}$$

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

$$\rho = \frac{P'}{P} \tag{3-3}$$

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 expanded pitch. Then the FFT is applied to the signal scaled at time axis.

$$\hat{S}(K) = \int_{-\infty}^{+\infty} \hat{s}(n) e^{-j\frac{n}{2\pi N}k} dn \tag{3-4}$$

This signal rewrite the real part and imaginary part.

$$\hat{S}(K)=Re[\hat{S}(K)]+jIm[\hat{S}(K)] \quad (3\text{-}5)$$

Where $Re[\hat{S}(K)]$ is the real part and $Im[\hat{S}(K)]$ is the imaginary part. And so, we use the definition of phase spectrum for Eq. (3-5) as

$$\phi(K)=\tan^{-1}\{Im[\hat{S}(K)]/Re[\hat{S}(K)\} \quad (3\text{-}6)$$

Like this, the obtained phase component is the signal which preserved phase characteristics of excitation.

Ⅳ. Spectrum control in frequency domain

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

$$S(K)=\int_{-\infty}^{+\infty} s(n)e^{-j\frac{n}{2\pi N}k} \quad dn \quad (4\text{-}1)$$

$$M(K)=10\log S^2(K) \quad (4\text{-}2)$$

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. The increasing of the fundamental frequency's interval on the given speech signal means the decreasing of pitch period in time domain. Therefore, if we know the pitch, the interval of fundamental frequency's interval, K_0, would be obtained as

$$K_0=\text{size}/\text{pitch} \quad (4\text{-}3)$$

where size is the length of window.

The logarithm of product can express by the addition of terms, hence the formant is extracted by Eq. (4-4) on spectrum.

$$S^*(K)=\frac{1}{K_0}\sum_{i=-K_0/2}^{K_0/2-1} M(K-i) \quad (4\text{-}4)$$

Where the $S^*(K)$ represents formant through the lifter, and K_0 is a fundamental frequency. If the extracted formant by Eq. (4-4) is subtracted from M(K) as Eq. (4-5), the flattened harmonics spectrum could be separated.

$$S_P(K)=M(K)-S^*(K) \quad (4\text{-}5)$$

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.

$$\hat{S}_p(K)=S_P(K\times\rho^{-1})$$
$$(K=0, 1, 2, 3, \cdots, \text{size}-1) \quad (4\text{-}6)$$

In Eq. (4-6), 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 the interval of the fundamental frequency by ρ^{-1} for compressing pitch.

Fig. 4-1 illustrates an example of the case that is reduced the interval of harmonics (expanded the pitch) by the scaling harmonics of spectrum. The interval of a harmonics' fundamental frequency in Fig. 4-1(b) is diminished less than that of Fig. 4-1(a).

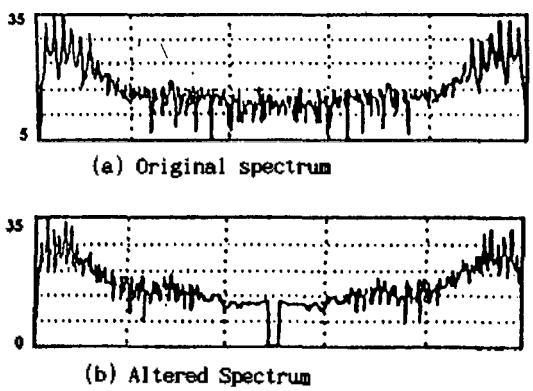

Fig. 4-1 An example to alter the pitch by harmonic scaling

V. Experimental and Results

For simulation, we used the 16-bit A/D converter interfaced with IBM-PC/486. The speech data composed of 3 Korean speaker's utterances(23 years old male and females) and following sentences were spoken each 3 times. The sampling frequency was 8kHz.

Utterance 1) / INSUNE KOMANUN CHUNJAE SONYUNUL JOAHANDA. /
Utterance 2) / KAMSAHAMNIDA. /
Utterance 3) / SOONGSILDAE JUNGBOTONGSIN KONGHAKKWA UMSEONG CHURI YUNGUSILIDA. /
Utterance 4) / JESUNIMKESEO CHUNJI CHANGJOWI KIOHUNWL MALSUMHASEOSSDA. /

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

The block diagram of the proposed algorithm shown in Fig.5-1. If the speech signal, $s(n)$, is performed pitch variation with arbitrary rate, that is, speech signal's pitch is adjusted by the method that compresses or expands the waveform of speech signal, it would make gross error on the spectrum by the variation of formant as well as pitch. Above all, to eliminate the influence of the upper formant and extract the phase component of excitation source, we filter the speech signal by the low pass filter with the pass band rate of one pitch, and scale it on the time axis, and obtain the phase component by FFT. After transformation the speech signal to frequency domain by FFT, we extract the magnitude component, make it passes through the lifter with the pass band rate of P^{-1}, and scale the difference between before and after the passing as scaling coefficient ρ^{-1}. It is obtained a pitch-varied magnitude spectrum by adding the harmonics scaled spectrum and the formant spectrum of the frequency domain. And then we obtained a pitch varied speech signal by the inverse fourier transform after adding it to phase component which is scaled and stored in

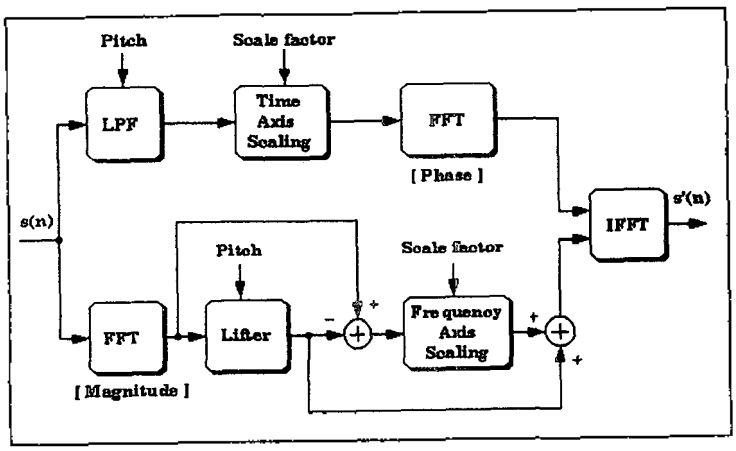

Fig. 5-1 The block diagram of proposed algorithm

time domain.

Fig. 5-2 illustrates the spectrum reduced a pitch of the speech waveform by 70%, and on the same speech waveform the spectrum is reduced by 70% after increasing to 140% shown as Fig. 5-3.

We measured the distortion of spectrum in accordance with the comparative criterion on the proposed pitch alteration method. The pitch changing is that force the pitch to increase or decrease. If we perform the pitch changing, the criterion of spectrum is eliminated. So, it is desired that the spectrum of the native(not changed) signal is the standard of evaluation of pitch alteration. Table 5-1 shows spectrum distortion rate in case of decreasing the pitch on utterance 1 by 50%, 55.56%, 62.5%, 71.43%, 83.33%, after increasing it by 200%, 180%, 160%, 140%, 120%.

If we expand the pitch by zero insertion and divide in two, the spectrum distortion is 8%. But in case of using the proposed algorithm, the spectrum distortion is decreased more 5.06% than that of existing pitch alteration method in time domain.

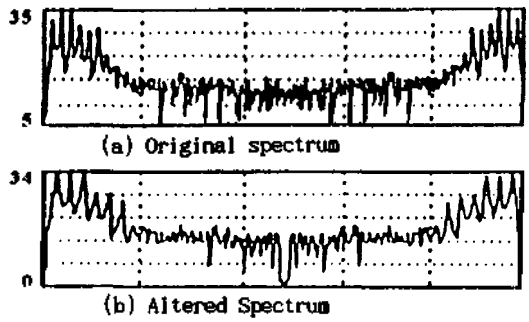

Fig. 5-2 A spectrum compressed the pitch as 70%
(a) original spectrum
(b) spectrum compressed the pitch as 70%

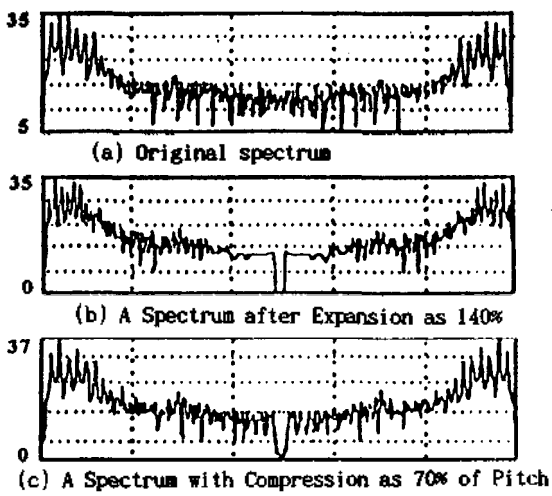

Fig. 5-3 A spectrum with compression as 70% of pitch after expansion as 140% of pitch
(a) original spectrum
(b) spectrum expanded the pitch as 140%
(c) spectrum with compression as 70% of pitch after expansion as 140%

Table 5-1. The spectrum distortion measurement

pitch alteration rate (%)	distortion rate (%)
200 ⇒ 50	2.94
180 ⇒ 55.56	2.71
160 ⇒ 62.5	2.16
140 ⇒ 71.43	2.28
120 ⇒ 83.33	1.72

VI. Conclusion

The purpose of this paper is to supply the basic technique for speech synthesis represented various individuality of speaker's by the alteration of pitch which represents the individuality of speaker's.

Thus, We proposed the pitch alteration for using the synthesis technique which supplys high quality speech in automatic response system without the limitation of the data quantity. In time domain, the speech signal is passed through the low pass filter(LPF) with a cut-off bandwidth as a pitch period. After the signal is scaled at time axis, the result is converted by Fourier transform for extraction phase component. Another, We convert a speech to frequency domain for pitch alteration, and separate the amplitude component. We force it to pass through the lifter, and scale the difference between before and after spectrum. By adding the scaled spectrum to the

amplitude spectrum in frequency domain. We obtain a pitch changed waveform. We add the amplitude component to the phase component which is stored in time domain, transform it by IFFT, obtain a pitch changed waveform. As the proposed method is performed in time-frequency domain, even if the pitch is changed, the phase information of a waveform is preserved. So the edition of waveform is simple in the synthesis based on the unit of a pitch period and the coupling between phonemes is natural for a little variation of the level. Besides, the information of filter is maintained for the unchanged formant component. So, the information of the opinion is saved successfully.

In result, spectrum distortion is above 8% in time domain pitch alteration, but we can reduced it below 3% by the proposed algorithm.

Reference

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
2. S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing*, Academic Press, 1985.
3. P. E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice Hall, 1987.
4. B. E. Caspers and B. S. Atal, "Changing pitch and duration in LPC synthesised speech using multipulse excitation," J. Acoust. Soc. Amer., suppl., vol. 73, no.1, pp.55, Spring, 1983.
5. M. G. Stella and F. J. Charpentier, "Diphon synthesis using multiples coding and a phase vocoder," in Proc. IEEE ICASSP'85, pp.740-744, 1985.
6. M. J. Bae, M. S. Lee, H. G. Lee, S. G. Ann "켑스트럼 분석에 의한 음성파형 코딩의 피치변경에 관한 연구," 제4회 신호처리 합동 학술대회 논문집, 제 4권 1호, pp.304-309, 1991년 9월.
7. 강동규, 김율제, 배명진, 안수길, "음성합성의 halving 기법에 의한 파형 코딩의 피치변경에 관한 연구," 한국음향학회 추계 발표회(국제 음향학회 논문집), pp.107-111. 1990년 11월 10일.
10. 배명진, "고음질 합성을 위한 피치변경법", 한국음향학회지, vol.12, no.2, pp.66-77.
11. B. E. Caspers and B. S. Atal, "Changing pitch and duration in LPC synthesised speech using multipulse excitation," J. Acoust. Soc., Amer., suppl., vol.73, no.1, pp.55, Spring, 1993.
12. J. D. Markel and A. H. ray, jr., *Linear Prediction of Speech Signals*, Springer-Vertag, 1976.
13. 유건수, 이성우, 배명진, 안수길, "고조파 스케일링에 의한 주파수 영역 피치 변경에 관한 연구," 대한 전자공학회 하계 종합학술대회 논문집, pp.637-640, 1992년 6월.
14. N. S. Jayant and P. Noll, *Digital Coding of Waveforms Principles and Applications to Speech and Video*, Prentice-Hall, Inc., 1978.
15. Shuzo Saito and Kazuo Nakata, *Fundamentals of Speech Signal Processing*, Academic Press, 1985.
16. John R. Deller, Jr., Jhon G. Proakis and John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, 1993.

▲Myungjin Bae : Vol. 13, No.1E 참고.