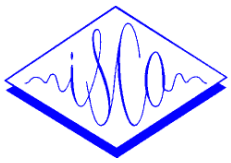

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Speech Analysis Technique for PSOLA Synthesis based on Complex Cepstrum Analysis and Residual Excitation

Yukio Mitome

NEC Corporation, 4-1-1 Miyazaki, Miyamae-ku, Kawasaki,
216 Japan

Abstract

This paper presents a new speech analysis method for waveform segment concatenation synthesis or PSOLA (Pitch Synchronous Over-Lap and Add) synthesis.

In the proposed method, two techniques are employed: a new algorithm for calculating the complex cepstrum and a technique for extracting waveform segments from human speech. Although the complex cepstrum analysis can estimate not only spectral envelope but also phase characteristics, every usual algorithm has its own problem in the application to voiced speech analysis. The proposed algorithm can solve these problems, and the combination of this technique with a residual excitation technique can extract waveform segments as accurate approximation of original human speech.

Evaluation tests for the proposed method was carried out, i.e. analysis-synthesis experiment and pitch modification synthesis. In the subjective listening tests for both synthesized speech and pitch modified speech, the listeners could not recognize the distortions. This shows the effectiveness of the proposed method.

1. Introduction

A waveform segment concatenation method is one of the most effective approaches to some kind of speech synthesis, such as pitch modification synthesis, time scale modification synthesis or text-to-speech synthesis[1]-[6]. Here, the waveform segment means the waveform whose duration is nearly equal to the pitch period or a few times as long as the pitch period.

This type of speech synthesis has two major merits. The first one is high intelligibility of synthesized speech and the other merit is a small computational cost in the synthesis process. The high intelligibility is most effectiveness of this synthesis method compared with parametric speech synthesis, such as LPC (Linear Predictive Coding) or formant synthesis. If these waveform segments are extracted beforehand from each pitch period of the human speech, and the duration of each waveform segment is longer than the pitch period, only addition is required at the overlapped portion. Because of this characteristic process, this kind of speech synthesis is called PSOLA(Pitch Synchronous Over-Lap and Add) or PSOLA-like synthesis.

The synthesized speech quality of this method depends on an analysis method that extracts waveform segments from human speech. Various analysis methods have been proposed for waveform synthesis, but they have their own problems.

In this paper, we first discuss these conventional analysis methods and their problems, then we chose the complex cepstrum as a basic approach to waveform segment extraction. Next, we discuss the problems of conventional algorithms for complex cepstrum, and then we propose a new algorithm to solve them. Moreover, a new complex cepstrum based method for waveform segment extraction is presented. Finally, subjective evaluation tests, which shows the effectiveness of the proposed method, are described.

2. Speech Analysis Methods

Various analysis methods for extracting waveform segments have been proposed, zero phase residual excited LPC[1], zero phase approximation of spectral envelope[3],[4], FFT (Fast Fourier Transform) based PSOLA, namely FD(Frequency Domain)-PSOLA, TD(Time Domain)-PSOLA[5],[6], and so on.

Among these methods, the types that equalize the phase are undesirable for high quality speech synthesis. Because phase equalization of a speech waveform or a residual waveform spoils the naturalness of speech quality. In other words, the synthesized speech sounds artificial. This shows that the accurate estimation of a phase is as important as the spectral envelope.

Although FFT based analysis or TD-PSOLA can approximate both spectral envelope and phase, separation of time varying characteristics of each pitch period is not sufficient. A residual excitation technique for LPC has the ability to separate time varying characteristics of each pitch period, because the energy in a residual wave is concentrated at the pitch period[7],[8]. However, in case of some phonemes, such as nasals, energy concentration of LPC residual is not sufficient.

According these considerations, the problems of speech analysis can be summarized as follows. The waveform segment extracted at each pitch period can be interpreted as time varying FIR(Finite Impulse Response) representation of speech production. From this view point, the impulse response at the beginning and the end should be dumped enough not to cause the discontinuity of synthesized speech. Moreover, its frequency characteristic must be an accurate approximation of the original speech signal.

The complex cepstrum can estimate not only the spectral envelope but also the phase characteristics[9],[10]. Moreover, homomorphic deconvolution can separate pitch period structure. Therefore, this analysis seems to be suitable for extracting waveform segments. However, the preliminary analysis-synthesis experiment showed that the synthesized speech did not have enough quality. We inferred that the major reasons are miscalculation of phase due to conventional algorithm for complex cepstrum and insufficient separation of time varying characteristics of pitch due to windowing.

3. Complex Cepstrum Analysis

In this section, the problems of conventional algorithms for complex cepstrum are discussed in detail and a new algorithm to solve them will be proposed.

3.1 Conventional Algorithms

The conventional algorithms for complex cepstrum have significant problems in voiced speech application. In this section, we discuss three conventional algorithms and their problems: a phase unwrapping method, a log derivative method, and a factorization method.

(a) Phase Unwrapping[9], [11]

The complex cepstrum is defined as an inverse Fourier Transform of a complex logarithm of the spectrum. The imaginary part of the complex logarithm of the spectrum, which corresponds to the phase component of the spectrum, must be continuous function of frequency. However, one can only obtain the principal value of the phase at discrete frequency, when using the DFT-based algorithm and a practical complex log function. The principal value is between $-\pi$ and $+\pi$. Therefore, one must find the discontinuity of the principal value version of the phase and add the multiple of $2 \cdot \pi$ to the principal value. In the application of this method to speech analysis, a misjudgement of phase continuity sometimes occurs especially in application to voiced speech analysis. The reason is that the spectrum of voiced speech has rapid change between harmonics.

(b) Log Derivative[9]

The log derivative method is based on the relationship between derivative in frequency domain and its time domain function, and it does not require continuity judgment. However, an aliasing phenomenon in the time domain causes severe spectral distortion when one applies this algorithm to voiced speech analysis. The reason is that the derivative of a log spectrum has pulses at notches between harmonics, so that its inverse Fourier transform does not decrease for a long time.

(c) Factorization[12]

The factorization algorithm does not require phase unwrapping process. It uses a theoretical formula to calculate complex cepstrum from a root of z-transform. This method assumes that one can find all the roots of the z-transform polynomial. In application to speech analysis, the degree of this polynomial is higher than 100 or 200, and it is nearly impossible to solve roots of this quite high order polynomial.

3.2 New Algorithm for Complex Cepstrum

The new algorithm proposed here also uses the relationship between derivative of a spectrum and its inverse Fourier transform. However, unlike conventional log derivative method, this relation is applied only to a phase component.

The basic idea is as follows. A phase component of the speech spectrum is much smoother than a log magnitude spectrum, so that aliasing in time domain due to differentiation is less significant. Moreover, the phase component can be obtained independently from a log magnitude component, because of symmetrical property of Fourier transform. Therefore, we can apply the relationship between Fourier transform and spectrum derivative to calculation of the phase component.

Let x(n) be windowed speech signal and c(n) be desired complex cepstrum. The complex cepstrum is defined by following equations.

$$c(n) = F^{-1}\left[C(e^{j\omega})\right] \tag{1}$$

$$C(e^{j\omega}) = \log\left[X(e^{j\omega})\right] \tag{2}$$

$$\begin{aligned} X(e^{j\omega}) &= F[x(n)] \\ &= \sum_{n=0}^{N-1} x(n) \cdot e^{-j\omega n} \end{aligned} \tag{3}$$

Where, $F[\bullet]$ and $F^{-1}[\bullet]$ denotes Fourier transform and Inverse Fourier transform respectively.

By differentiating eq.(3) we obtain the general relationship between the signal x and the differential of its Fourier transform X.

$$j X'(e^{j\omega}) = F[n \cdot x(n)] \tag{4}$$

$$x(n) = F^{-1}\left[j X'(e^{j\omega})\right] / n \tag{5}$$

We can describe the log spectrum in terms of real and imaginary part, that is log magnitude and phase components.

$$\begin{aligned} C(e^{j\omega}) &= \log\left[X(e^{j\omega})\right] \\ &= A(e^{j\omega}) + j\Theta(e^{j\omega}) \end{aligned} \tag{6}$$

Note that, A is even function of frequency, and Θ is odd function of frequency, because the original signal x(n) is real.

We define a(n) and b(n) as

$$a(n) = F^{-1}\left[A(e^{j\omega})\right] \tag{7}$$

$$b(n) = F^{-1}\left[j\Theta(e^{j\omega})\right] \tag{8}$$

By using the linearity principle of the Fourier transform, and real-imaginary separation of A and Θ, the complex cepstrum can be expressed as

$$c(n) = a(n) + b(n) \tag{9}$$

According to eq. (7), the term a(n) in eq. (9) is normal cepstrum, which can be easily calculated using a real log function and a DFT algorithm. Therefore, we only need to obtain b(n).

Applying eq. (5) to eq. (8), we obtain

$$\begin{aligned} b(n) &= F^{-1}\left[j \cdot \left(j\Theta(e^{j\omega})\right)' \right] / n \\ &= F^{-1}\left[-\Theta'(e^{j\omega})\right] / n \end{aligned} \tag{10}$$

By differentiating eq. (2) and eq. (6), we obtain two equations for the derivative of log spectrum.

$$C'(e^{j\omega}) = \frac{X'(e^{j\omega})}{X(e^{j\omega})} \tag{11}$$

$$C'(e^{j\omega}) = A'(e^{j\omega}) + j\Theta'(e^{j\omega}) \tag{12}$$

Multiply j to eq. (12) becomes

$$j C'(e^{j\omega}) = j A'(e^{j\omega}) - \Theta'(e^{j\omega}) \tag{13}$$

From eq. (11) and (13), we obtain

$$-\Theta'(e^{j\omega}) = \mathrm{Re}\left[\frac{jX'(e^{j\omega})}{X(e^{j\omega})}\right] = \mathrm{Re}\left[F[n\cdot x(n)]/F[x(n)]\right] \quad (14)$$

By substituting eq.(14) for eq.(10), we obtain a theoretical formula for b(n).

$$b(n) = F^{-1}\left[\mathrm{Re}\left[F[n\cdot x(n)]/F[x(n)]\right]\right]/n \quad (15)$$

The practical computation algorithm based on this principle can be summarized as follows.

(a) Calculate X(k) = DFT[x(n)] and Y(k) = DFT[n·x(n)].
(b) Calculate p(n) = IDFT[Re[Y(k) / X(k)]] and b(n) = p(n)/n.
(c) Calculate a(n) = IDFT[log | X(k) |].
(d) Calculate c(n) = a(n) + b(n).

Here, DFT[ ], IDFT[ ], Re[ ] and | | means Discrete Fourier transform, Inverse DFT, real part of complex number, and absolute value of complex number, respectively.

Fig.1 shows the blockdiagram of this algorithm.

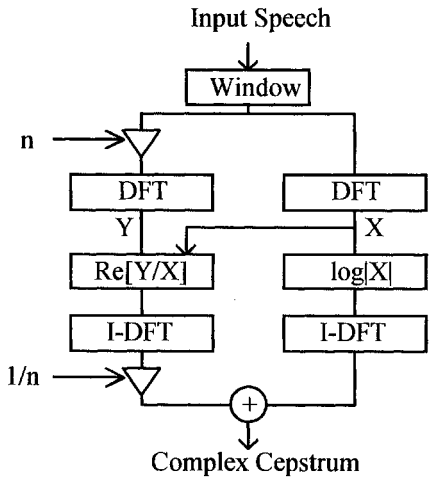

Fig. 1. A New Algorithm for Complex Cepstrum.

This algorithm has following features. No logical decision is required, so that misjudgement never occurs. It is a DFT-based algorithm with finite procedure, and consequently assures getting the solution.

4. Waveform Segment Extraction

Although the algorithm for complex cepstrum proposed in previous section can estimate accurate phase as well as spectral envelope, the problem of pitch separation is not yet solved. In this section, a new method is proposed.

4.1 A Complex Cepstrum based Residual Excited FIR Method

The proposed method for waveform segment preparation is based on complex cepstrum analysis for spectral estimation and residual excitation technique for separation of time varying characteristics. As we discussed in previous section, a RELP(residual excited LPC) method has the ability to separate time varying characteristics of each pitch period, but energy concentration of LPC residual is not sufficient in case of some phonemes. Complex cepstrum analysis can estimate both spectral envelope and the phase characteristics, so that a residual wave has strong concentration of its power at each pitch period.

Fig.2 shows the outline of this speech analysis method, and the following is the procedure of this method.

At first, complex cepstrum is calculated using the proposed algorithm. After liftering this complex cepstrum, an FIR filter coefficients are calculated as "the First Stage Approximation of Speech." This FIR filter is the same as the final approximation in the conventional homomorphic deconvolution. Next, the inverse filtering produces a residual waveform, from which a segment corresponding one pitch period is extracted as source signal for the first stage FIR. Finally, the output signal from the FIR filter is obtained as "the Second Stage Approximation."

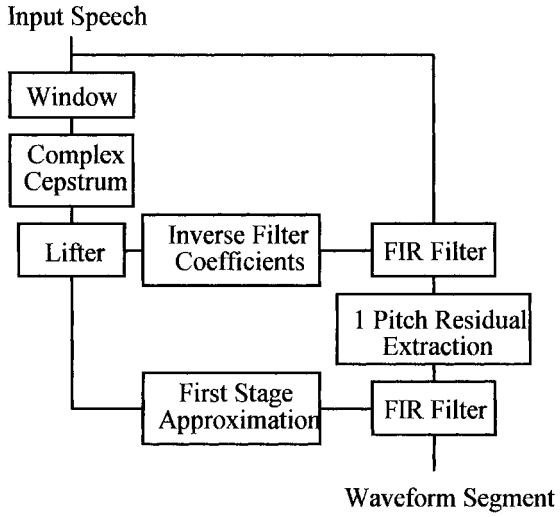

Fig. 2. An Outline of Waveform Segment Extraction.

4.2 Lifter Consideration

Several kinds of lifters have been proposed, such as "Low-Pass," "Notch," and "Comb." The notch lifter cuts the region including pitch component, and the comb lifter cuts several regions including pitch period and the multiple of pitch period.

These two lifters pass the high quefrency components so as to obtain a close estimation to the harmonics envelope. However, we should not use high quefrency components in a DFT- based algorithm, because these components suffered the aliasing due to spectrum sampling in frequency domain.

To obtain more accurate estimation of harmonics envelope than a lowpass lifter, we propose a new lifter named "Aliasing Lifter." If the spectrum is sampled at each harmonics frequency, aliasing which is synchronized with pitch period will occur. Unlike aliasing in high quefrency components due to DFT, this aliasing is desirable, because this is synchronous with pitch and the extracted envelope fits the spectral harmonics. This process can be done as follows.

$$\tilde{c}(n) = c(n) + \sum_k \left\{c(k\cdot N_p + n) + c(k\cdot N_p - n)\right\} \quad (16)$$

To avoid aliasing due to DFT, only components around the pitch period or several repetitions are used.

4.3 Inverse Filter Implementation

An LPC filter assumes that the speech production can be modeled by a minimum phase filter, so that the inverse filter is realized using all zero filter whose coefficients are the same as an LPC synthesis filter.

Unlike the LPC filter, the first stage FIR filter does not assume minimum phase, so that zeros can exist outside the unit circle of z-plane. Therefore, the stability of the theoretical inverse system is not assured. However, an FIR filter using windowing technique can realize the inverse filter. Although FIR implementation does not assure the ideal inverse system, it is practically sufficient to obtain a residual waveform.

5. Evaluation

Two evaluation tests for the proposed method have been carried out. The first one is analysis-synthesis and the other is pitch modification synthesis.

The speech material consists of 12 sentences, which was uttered by a professional announcer in a low noise recording booth. The total duration of this material is about one minute. The data was digitized at 11 kHz, and the pitch was marked semiautomatically.

The voiced waveform segments were extracted using the proposed method, while the unvoiced segments were extracted by tapering at the end. In the pitch modification synthesis, pitch periods are changed by multiplying coefficient to them.

In the subjective listening tests of both synthesized speech and pitch modified speech, the listeners could not recognize the distortions. This shows the effectiveness of the proposed method.

6. Conclusion

A new speech analysis method for waveform concatenation (PSOLA or PSOLA-like) synthesis was proposed. This method consists of two techniques, a new algorithm for complex cepstrum and a waveform segment extraction method.

The proposed algorithm for complex cepstrum does not require logical decision nor unlimited repetition for solving equation, and the obtained complex cepstrum is more accurate than that obtained using the conventional method.

In the subjective listening tests for both synthesized speech and pitch modified speech, the listeners could not recognize the distortions. This shows the effectiveness of the proposed method.

The proposed method was applied to Japanese text-to-speech synthesis[14], and the system could synthesize clear and natural voice.

References

[1] Fushikida, K. and Ochiai, K.,"An Automatic Analysis-Synthesis System for Speech Segment Compilation Synthesis," Tras. of the Committee on Speech Research, the Acoustical Society of Japan, S74-23, Nov.1974(in Japanese)

[2] Mitome, Y. and Fushikida, K.,"A Speech Synthesis Method for Unrestricted Words using Pitch Synchronous Interpolation between CV,VC Waveforms," Proc. of ASJ Meeting, 1-7-2, May 1981 (in Japanese)

[3] Miki, K., Yazu, T., Morito, M., and Yamada, K., "Speech Synthesis by Symmetric Segments," Proc. of ASJ Meeting, 1-2-16, Oct. 1983 (in Japanese)

[4] Yazu,T., Yamada,K., "The Speech Synthesis System for an Unlimited Japanese Vocabulary," Proc. ICASSP, 1986

[5] Charpentier, F. and Moulines, E.,"Text-to-Speech Algorithms based on FFT Synthesis," Proc. of ICASSP '88, pp. 667-670, 1988

[6] Moulines, E. and Charpentier, F.,"Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones," Speech Communication 9, pp.453-467, 1990

[7] Iwata, k., Ichiwaka, M., Ozawa, K. and Watanabe, T., "Speech Synthesis System using Pitch Controled Residual Wave Excitation," Proc. of ASJ Meeting, 3-2-7, Oct. 1988 (in Japanese)

[8] Iwata, K., Mitome, Y., Kametani, J., Akamatsu, M., Tomotake, S., Ozawa, K. and Watanabe, T., "A Rule Based Speech Synthesizer using Pitch Controlled residual Wave Excitation Method," Proc. ICSLP '90, pp.185-188, 1990

[9] Oppenheim, A.V. and Schafer, R.W., Digital Signal Processing, Chap.10

[10] Seiyama, N., Tkagi, T., Umeda, T. and Miyasaka, E., "A High Quality Pitch Modification Method by Complex Cepstrum Analysis-Synthesis," IEICE Technical Report, SP90-28,1990 (in Japanese)

[11] McGowan, R. and Kuc, R., "A Direct Relation Between a Signal Time Series and Its Unwrapped Phase," IEEE Trans. ASSP-30, No.5, Oct. 1982.

[12] Tribolet, J. M., "A New Phase Unwrapping Algorithm," IEEE Trans. ASSP-25, No.2, Apr. 1977.

[13] Steiglitz, K. and Dickinson, B.,"Phase Unwrapping by Factorization," IEEE Trans. ASSP-30, No.6, Dec. 1982.

[14] Takahashi, K., Iwata, K., Nagano, K and Mitome, Y.,"Japanese Text-to-Speech Conversion Software for Personal Computers," ICSLP '94, 29.1, Sep. 1994