EFFICIENT REPRESENTATION OF SHORT-TIME PHASE BASED ON GROUP DELAY

Hideki BANNO[1] *Jinlin LU*[1] *Satoshi NAKAMURA*[1] *Kiyohiro SHIKANO*[1] *Hideki KAWAHARA*[2]

[1]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara 630-01, Japan

[2]Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama-shi, Wakayama 640, Japan

E-mail: hideki-b@is.aist-nara.ac.jp

ABSTRACT

An efficient representation of short-time phase characteristics of speech sounds is proposed, based on recent findings which suggest perceptual importance of phase characteristics. Subjective tests indicated that the synthesized speech sounds by the proposed method are indistinguishable from the original speech sounds with a moderate data compression. The proposed representation uses lower-order coefficients of inverse Fourier transform of the group delay of speech. It also alleviates the voiced/unvoiced decision, which is an indispensable part in conventional speech coding algorithms. These features make our method potentially very useful in many applications like speech morphing.

1. INTRODUCTION

The conventional synthesized speech has been reconstructed with zero phase or minimum phase in analysis-synthesis methods based on the fact that the phase information is not believed so important for auditory perception. Honda et al.[1] have proposed a speech coder using phase equalization, and showed its effectiveness. The synthesized speech by these methods suffers from the characteristic sound so called 'buzziness'. Another problem, which almost all analysis-synthesis systems suffer from, is a difficulty of the decision of voiced or unvoiced frames, so several methods are proposed to solve these problems.

In speech coding, one of the most general methods is to encode a residual of LPC[2]. The LPC residual includes amplitude information which can not be represented using LPC coefficients and phase information of the input speech. So the quality of the LPC residual-based synthesized speech is much better than the quality of the zero phase or minimum phase synthesis. In addition, there is no need of voiced/unvoiced decision with the LPC residual-based method . The LPC residual-based method is widely used in speech coding because the bit rate can be reduced. However, it is known that the bit rate reduction less than a certain value leads to degradation in speech quality. It is also difficult to apply to speech modification, such as speech transformation and speech morphing. Though the other methods are improved in order to reduce buzziness of zero phase synthesis[3, 4], they still have similar problems.

A new speech manipulation method, STRAIGHT, proposed by one of the author [5] provided another evidence that phase (or group delay) information plays an important role in high quality speech synthesis. STRAIGHT is essentially a channel VOCODER using minimum phase impulse responses to resynthesize speech. However, each impulse response is modified using an all-pass filter with randomly manipulated group delay in higher ($>$ 2kHz, typically) frequency region. This manipulation of group delay found to add 'naturalness'. Even though STRAIGHT provides high quality synthetic speech sounds, it is still possible to discriminate the difference between the original speech and the synthetic speech, especially in 8kHz sampling rate. STRAIGHT also has the similar voiced/unvoiced decision problem as the other analysis-synthesis methods do.

In a study of auditory perception, Patterson proves that the phase information is important for auditory perception. He also proposes a model which simulates our auditory perception properly[6]. As this model takes no consideration for synthesis, it can not be applied to applications such as speech coding and speech transformation.

We propose a new representation of the short-time phase for applying to speech coding or speech transformation. A representation of short-time phase is required to keep important parts for auditory perception, to reduce the bit rate, and to interpolate between parameters for speech transformation and speech morphing. The proposed method has a large possibility to satisfy these requirements.

Generally the range of phase spectra which are obtained by discrete Fourier transform is limited between $-\pi$ and π, so the phase spectra have discontinuity. To avoid this discontinuity caused by the limitation, a group delay is used in our method. It may be also possible to interpolate between phases by using the group delay, because a group delay is correspond to delay in time domain[7].

2. GROUP DELAY

In this section, we discuss how to estimate the group delay.

2.1. Group Delay Estimation

Let $x(t)$ is an input signal, then its Fourier transform $X(\omega)$ is given by

$$X(\omega) = \mathcal{F}[x(t)] = a(\omega) + jb(\omega) \tag{1}$$

where $F[\cdot]$ represents Fourier transform. Its amplitude and phase can be written as

$$|X(\omega)| = \sqrt{a^2(\omega) + b^2(\omega)} \tag{2}$$

$$\theta(\omega) = \arctan\left(\frac{b(\omega)}{a(\omega)}\right) \tag{3}$$

A group delay is defined by

$$D(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{4}$$

The equation of phase definition (3) becomes

$$\tan\theta(\omega) = \frac{b(\omega)}{a(\omega)} \tag{5}$$

By the differential calculus of (3), the group delay can be rewritten as

$$-\frac{d\theta(\omega)}{d\omega} = \frac{-\frac{db(\omega)}{d\omega}a(\omega) + b(\omega)\frac{da(\omega)}{d\omega}}{a^2(\omega) + b^2(\omega)} \tag{6}$$

By the property of Fourier transform $(-jt)x(t) \leftrightarrow \frac{dX(\omega)}{d\omega}$, we get (7)

$$\frac{dX(\omega)}{d\omega} = \frac{da(\omega)}{d\omega} + j\frac{db(\omega)}{d\omega} = \mathcal{F}[-jtx(t)] \tag{7}$$

Then, the group delay $D(\omega)$ is expressed as

$$D(\omega) = \frac{-\mathrm{Im}[\mathcal{F}[-jtx(t)]]a(\omega) + b(\omega)\mathrm{Re}[\mathcal{F}[-jtx(t)]]}{a^2(\omega) + b^2(\omega)} \tag{8}$$

We have to decide a reference point of input speech estimating the group delay. The reference point is the peak of every pitch waveform in input speech in our experiment. We call the reference point 'pitch mark'. A one-pitch waveform is estimated with windowing the input signal according to the pitch mark. The one-pitch waveform is twice as long as an interval of the pitch mark. If there is no mark at a certain frame, we consider the frame unvoiced, and the frame is windowed with a constant length.

2.2. Synthesis Using Group Delay

The phase spectrum can be obtained by integration of the group delay.

$$\theta'(\omega) = -\int_0^{\omega} D(\omega)d\omega \tag{9}$$

Practically, the integration is approximated with the cumulative summation.

$$\theta'(k) = -\frac{2\pi}{N}\sum_0^k D(k) \tag{10}$$

Here, N is the number of FFT points.

The complex spectrum is obtained using both the phase spectrum and the amplitude spectrum. It leads to the one-pitch waveform by inverse Fourier transform.

$$X'(\omega) = |X(\omega)|e^{j\theta'(\omega)} \tag{11}$$

$$x'(t) = \mathcal{F}^{-1}[X'(\omega)] \tag{12}$$

The synthesized speech is reconstructed using this one-pitch waveform by the overlap-add method. In this algorithm, even in the unvoiced frame, it is not required to convolute a random noise source as in the zero phase synthesis.

The above algorithm makes it possible to synthesize high quality speech. Occasionally, the synthesized speech suffers from click noise caused by the summation approximation error of (10). We explain the way of dealing with the click noise in the next section.

2.3. Click Noise Removal

2.3.1. Pitch Marking Strategy

It is known that phase spectral fluctuation becomes unstable in some pitch mark positions[8]. In the large phase fluctuation, the above algorithm has a possibility of resulting large error by the summation of the group delay. In case that the input signal is a periodic impulse, the pitch mark is positioned at the peak of an impulse, while phase spectral fluctuation becomes smallest[8]. Based on this idea, we can say the error in the summation (10) is reduced if pitch mark is put on the suitable position. By a preliminary experiment, the best result was produced when we choose the position as the peak of the low pass filtered (about < 500Hz) speech.

2.3.2. Limitation of Phase

The real phase spectrum has limitation that the value in radian frequency 2π is $2n\pi$ (n: integer). The estimated phase spectrum does not satisfy the requirement because of approximation of the integration. So we give a linear phase component to the estimated phase for satisfying the limitation.

2.3.3. FFT Points

Increasing FFT points leads to improve approximation accuracy of the integration, and to decrease the error of waveform. In our experiment, the number of the FFT points is 8192, slightly larger than in normal speech analysis.

These error reduction algorithms make it possible to synthesize high quality speech which we can hardly distinguish from original speech.

3. GROUP DELAY INFORMATION COMPRESSION

In the previous section, we show that it is possible to synthesize high quality speech using the group delay. However a large amount of data, a half of FFT points (because the group delay is even function), are required. In this section, we propose a method to compress the group delay information, and we show that our compression method represents the phase information efficiently.

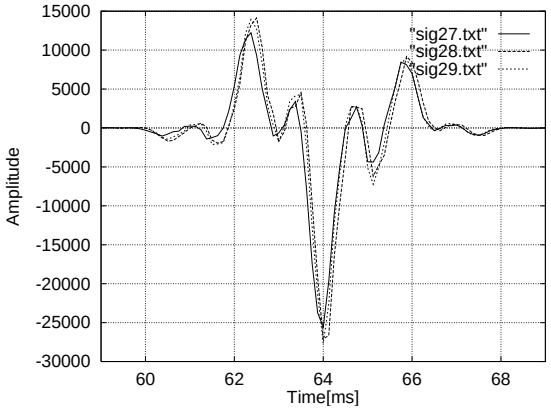

Figure 1: Example of one-pitch waveform

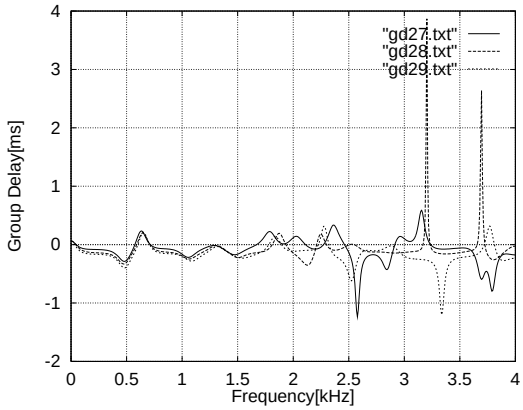

Figure 2: Example of group delay

3.1. Characteristics of Group Delay

Figure 2 shows the estimated group delay from time series of one-pitch waveform in Figure 1.

We can find the following characteristics, from Figure 2.

1. The group delay values change smoothly in time domain, while the waveform changes little.
2. The group delay values change smoothly in frequency domain except for a few peaks in the group delay.
3. The detailed group delay values have no correlation with the waveform shape, while the global group delay values are related to the waveform shape.

These characteristics imply that the detailed group delay values may be not so important for auditory perception. From this implication, we assume that the global shape of the group delay is important in speech. In other words, we can compress the phase information without degradation, when the global shape of the group delay is used. We explain how to estimate the global shape of the group delay in next subsection.

3.2. Time Domain Smoothed Group Delay

The smoothed group delay contains only the global information. we use a low pass filter for smoothing the group delay. We can extract the global shape of the group delay by low-order coefficients of Fourier transformed group delay. Fourier transformation and inverse Fourier transformation is equivalent because the group delay is even function. Then we substitute Fourier transform with inverse Fourier transform. The inverse Fourier transformed group delay coefficients become the parameter in time domain. We call the coefficients TSGD (Time-domain Smoothed Group Delay). TSGD is expressed as

$$d(n) = \mathcal{F}^{-1}[D(k)] \qquad (0 \leq n \leq \alpha) \qquad (13)$$

Here, $D(k)$ is the group delay in discrete time domain, α is the order for windowing. The order of the TSGD coefficients is independent of the FFT points because TSGD is time domain.

If you want to get synthesized speech using TSGD, estimate the group delay by calculating Fourier transform of TSGD, utilize the synthesis algorithm from the group delay as we mentioned in Section 2.

Table 1: Analysis conditions

Sampling Frequency	8kHz
Window	Gaussian Window
Window Length	Variable with Pitch
FFT point	8192 points

4. EVALUATION EXPERIMENT

4.1. Experiment 1: Objective Evaluation

In this section, we evaluate the performance of TSGD objectively using the segmental SNR, in order to investigate relation between the coefficient order of TSGD and the waveform shape.

We compute the segmental SNR for each synthesized speech in several coefficient orders of TSGD. In our experiment, the SNR is not compared with real speech but with only windowed synthesized speech. As Gaussian window is good for estimating the group delay by a preliminary experiment, we use it. Table 1 shows the experimental conditions. Speech materials are two short sentences which are uttered by two male and two female (8 utterances in total).

Figure 3 shows the relation between the TSGD coefficient order and the segmental SNR. From this figure, we can see the segmental SNR has almost saturated value over 100 order of TSGD. This imply that the waveform information is concentrated below the 100th coefficient of TSGD. That is to say that waveform information exists in the global shape of the group delay.

The segmental SNR value using lower 20 coefficients of TSGD is over 20dB. It is good enough to reconstruct waveform. In next section, we discuss if the segmental SNR value in less than 20 TSGD coefficients is sufficient for auditory perception.

4.2. Experiment 2: Subjective Evaluation

Subjective evaluation is carried out by 6 subjects. The subjects evaluate the speech quality in the 5 grades. We

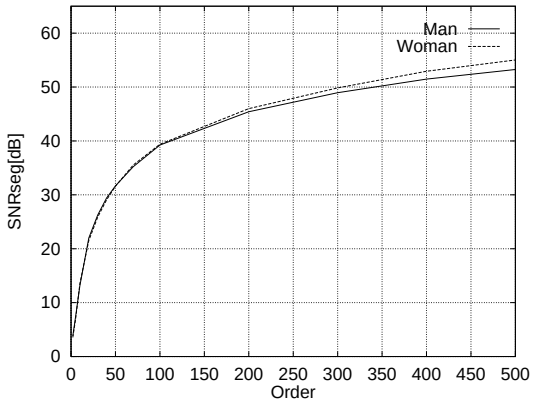

Figure 3: Relation between TSGD order and segmental SNR

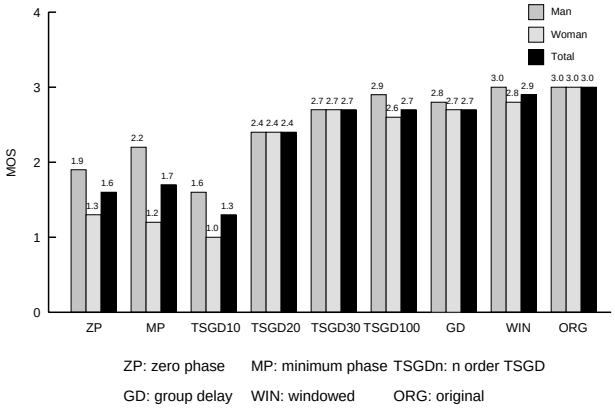

Figure 4: Results of subjective test by opinion test

instruct them to pay attention to speech timbre. The analysis conditions are the same as in Experiment 1.

Speech materials are original speech (ORG), synthesized speech only by windowing (WIN), synthesized speech using the group delay (GD), synthesized speech with TSGD, synthesized speech with zero phase (ZP), and synthesized speech with minimum phase (MP).

The zero phase and minimum phase speech are synthesized in order to avoid degradation as much as possible, as follows:

1. Estimate the detailed fundamental frequency using the method in [9].
2. Window the speech from pitch mark information in order to get one-pitch waveform.
3. The windowed one-pitch waveform is transformed to zero phase or minimum phase with FFT.
4. Convolute random noise source in unvoiced frames.
5. Using estimated fundamental frequency, arrange the one-pitch waveform. Control fundamental frequency more accurately with the method in [5].

By applying the above processing, the difference between zero or minimum phase speech and original speech is only caused by short-time phase differences.

Figure 4 shows the mean opinion scores of the subjective test. The zero and minimum phase result in lower scores. In contrast, high quality speech equivalent to original speech can be synthesized, when more than low 30 TSGD coefficients are used. As the TSGD coefficient order is decreased than low 30 TSGD coefficients, the quality of the synthesized speech becomes lower.

The another merit of the proposed method is that the synthesized speech with TSGD does not need to convolute the random noise source at all. The needlessness of the random noise source convolution makes it possible to reduce degradation caused by the voiced/unvoiced decision error for applications such as speech coding and speech morphing.

5. CONCLUSION

The time domain smoothed group delay (TSGD) representation was proposed taking advantage of a psychological observation that the global shape of the speech group delay plays an important role in reproduction quality. It was also shown that the TSGD coefficients are efficient representation to preserve waveform information in terms of the segmental SNR measure.

A series of subjective tests demonstrated that the lower 30 coefficients of TSGD is sufficient to produce high quality speech sounds which are almost indistinguishable from the original speech sounds. It was also confirmed that simply using minimum phase or zero phase impulse response to resynthesize speech signals damages perceptual quality significantly.

The proposed method is useful for many applications such as speech coding, speech transformation and speech morphing because of easy implementation and robustness.

6. REFERENCES

[1] M.Honda. Speech Analysis-Synthesis Using Phase-Equalized Excitation. *Technical Report of IEICE*, SP89-124, pp.1–8, 1989 (in Japanese).

[2] M.R.Schroeder. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. *Proc. ICASSP 1985*, pp.937–940, 1985.

[3] I.M.Trancoso, R.G.Gomez ans J.M.Tribolet. A Study on Short-Time Phase and Multipulse LPC. *Proc. ICASSP 1984*, pp.10.3.1–10.3.4, 1984.

[4] P.Hedelin. Phase Compensation in All-Pole Speech Analysis. *Proc. ICASSP 1988*, pp.339–342, 1988.

[5] H.Kawahara and I.Masuda. Speech Representation and Transformation based on Adaptive Time-Frequency Interpolation. *Technical Report of IEICE*, EA96-28, pp.9–16, 1996 (in Japanese).

[6] Roy D. Patterson. The sound of a sinusoid: Time-interval models *J. Acoust. Soc. Amer.*, vol.96, No.3, pp.1419–1428, 1994.

[7] B.Boashash. Estimating ans Interpreting The Instantaneous Frequency of a Signal. *Proc. IEEE*, vol.80, No.4, pp.519–569, 1992.

[8] L.R.Rabiner and R.W.Shafer. Digital Processing of Speech Signal. *Prentice-Hall*, 1978.

[9] H.Kawahara and Alain de Chaveighné. Error Free F0 Extraction Method and Its Evaluation. *Technical Report of IEICE*, SP96-96, pp.9–18, 1997 (in Japanese).