SINUSOIDAL REPRESENTATION AND AUDITORY MODEL-BASED PARAMETRIC MATCHING AND SMOOTHING AND ITS APPLICATION IN SPEECH ANALYSIS/SYNTHESIS

Oscar C. Au, Wanggen Wan**, Cyan L. Keung, Chi H. Yim*

Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
Email: *eeau@ust.hk, **eewwg@ee.ust.hk

10.21437/Eurospeech.1999-498

ABSTRACT

This paper presents a parametric matching and smoothing method that is applied to a sinusoidal representation and auditory model-based speech analysis/synthesis system. A 2.6kbps speech- coding algorithm is finally derived based on the speech analysis/synthesis system. The synthetic speech is almost same as that of 3.25kbps speech coding algorithm with overlapping and adding method. A linear interpolation method is utilized to smooth the amplitude parameters, and a nonlinear polynomial interpolation method is used to smooth the frequency and phase parameters. The experimental results demonstrate that the parametric matching and smoothing method can reduce the bit-rate with the speech quality unchanged when it is applied to the sinusoidal representation and auditory model-based speech-coding algorithm.

1. INTRODUCTION

Speech is often divided into frames and analyzed frame by frame in digital speech processing, especially in the parametric coding of speech. Speech can be processed by frames is because it has a quite steady statistic property within a period of time, e.g. 20-30ms. This property brings a great convenience to speech analysis, but it makes trouble when synthesizing speech. There will be discontinuation for parameters between the frames, which will produce something like noise in the synthetic speech if without adopting any parametric smoothing. One simple method of parametric smoothing is “overlap and add” which is applied in reference [1][2]. This method requires that the interval between the frames is not too long, it is better to be within 10ms. That is to say, the speed of speech analysis should be higher than 100Hz, e.g. 100 frames per second. The larger the interval is, the worse the quality of synthetic speech is. This result is obviously not beneficial to the low bit-rate speech coding because high speed of speech analysis will raise the bit-rate of speech coding.

In this paper, speech is analyzed using sinusoidal representation and auditory model. A new auditory spectrum-based speech feature is used. In order to raise the quality of synthetic speech and reduce the bit-rate, parametric matching and smoothing are utilized and combined with the auditory model-based speech synthesis. We will first give a brief introduction for sinusoidal representation, and then introduce speech analysis based on auditory model. Parametric matching and smoothing method will be presented after introducing the extraction of speech feature. Finally the experimental results will be presented.

2. SPEECH ANALYSIS

2.1. Sinusoidal representation

It is well known that speech can be represented using the following formula:

$$s(t) = \int_0^t h(t-\tau;t)e(\tau)d\tau \qquad (1)$$

where e(t) is an excitation signal and h(τ;t) is time-variant impulse response function of vocal tract model. Instead of being classified to be periodical pulse or white noise for voiced or unvoiced speech, e(t) can be represented in the following form in the sinusoidal representation:

$$e(t) = \sum_{l=1}^{L(t)} a_l(t)\cos[\int_0^t \omega_l(\sigma)d\sigma + \phi_l] \qquad (2)$$

where $a_l(t)$, $\omega_l(t)$ and ϕ_l are the amplitude, frequency and initial phase of the lth component in the excitation signal.

Assume that h(τ;t) has the following transfer function:

$$H[\omega(t);t] = M[\omega(t);t]\exp\{j\Phi[\omega(t);t]\} \qquad (3)$$

where M[ω(t);t] and Φ[ω(t);t] are time-variant amplitude characteristics and phase characteristics. Combing Equation (1), Equation (2) and Equation (3), and using the properties of linear system, we can obtain:

$$s(t) = \sum_{l=1}^{L(t)} a_l(t) M[\omega_l(t);t] \cos\{\int_0^t \omega_l(\tau)d\tau + \Phi[\omega_l(t);t] + \phi_l\} \quad (4)$$

we can express Equation (4) in a more concise form:

$$s(t) = \sum_{l=1}^{L(t)} A_l(t) \cos\Psi_l(t) \quad (5)$$

where

$$A_l(t) = a_l(t) M[\omega_l(t);t] \quad (6)$$

$$\Psi_l(t) = \int_0^t \omega_l(\tau)d\tau + \Phi[\omega_l(t);t] + \phi_l \quad (7)$$

From Equation (5), we can find that speech can be represented using many sine (cosine) waves that have different amplitudes and different phases. What is important is how we can find those amplitudes and phases that can represent speech most efficiently. For the sake of low bit-rate speech coding, we hope to use few parameters to represent the speech more accurately. This can be done with the combination of sinusoidal representation and auditory model that will be discussed below.

2.2. **Auditory spectrum-based speech feature and its extraction**

In order to extract more efficient speech feature, an auditory model [1] is used with the combination of sinusoidal representation. Principles of the auditory model and speech feature extraction are presented in Figure 1 and Figure 2 respectively. In Figure 1 and Figure 2, SDCM stands for the second order difference cochlear model [3], and PANPM stands for primary auditory nerve processing model whose function is described in reference [1]. Figure 3 presents an amplitude spectrum for a frame of original speech, and Figure 4 shows the corresponding auditory spectrum-based speech feature. Those frequency components presented in Figure 4 will be used in Equation (5) to reconstruct speech.

3. PARAMETRIC MATCHING AND SMOOTHING

There are different number of auditory spectrum lines for different frames, and even the number of spectrum lines are sometimes same for some frames, they are not continuous and smooth between the frames. In order to raise the quality of synthetic speech, speech parameters should be matched and smoothed between the frames after the parameters decoding.

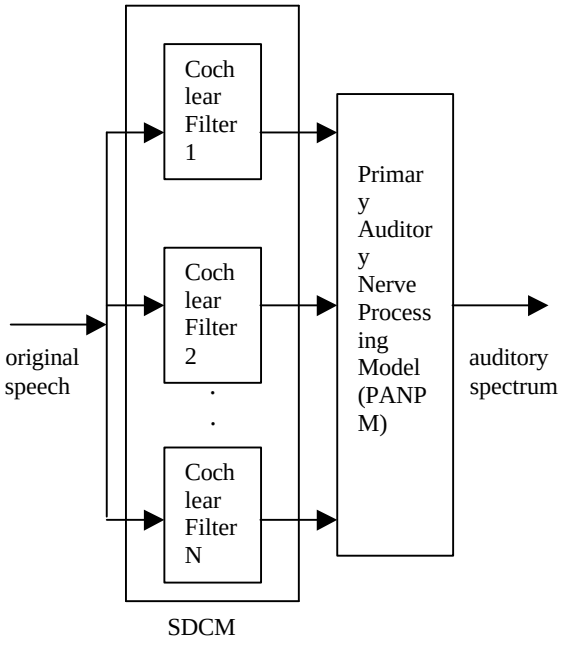

Figure 1. Structure of auditory model

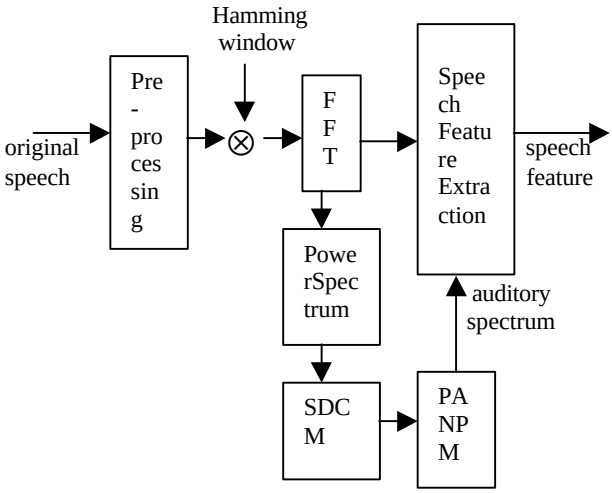

Figure 2. Principle of speech feature extraction

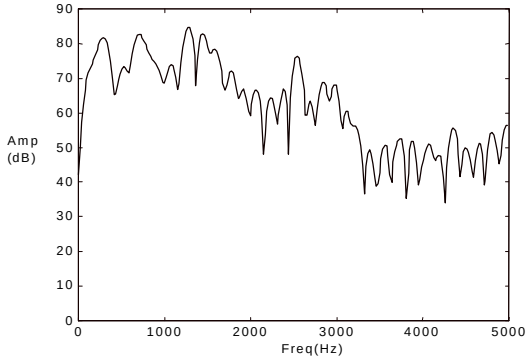

Figure 3. Spectrum of a frame of original speech

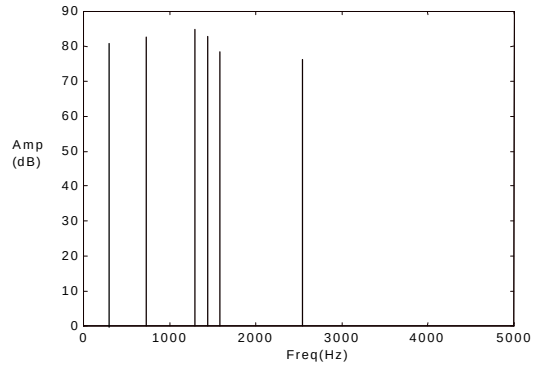

Figure 4. Auditory spectrum-based speech feature corresponding to Figure 3.

The principle of parameters matching is described below: if there is a frequency in the next frame(k+1th frame) whose difference between the frequency of current frame(kth frame) is smaller than Δ, the frequency track will continue and go into the next frame, otherwise the track will end in the next frame. The value of Δ is generally 10% of all the frequency range concerned. If there is a frequency in the next frame that can not find a matched frequency in the current frame, then a new frequency track will begin in the next frame. In such a way, frequency track can be considered to be continuous, but nerve smooth.

There are three kinds of speech parameters to be used, i.e. amplitude, frequency and phase where frequency and phase can be considered to be one kind of parameter, because the frequency is the first-order derivation of the phase. So frequency smoothing can also be considered to be phase smoothing. We use the following nonlinear polynomial to smooth the phase.

$$\tilde{\theta}_l(t) = a + bt + ct^2 + dt^3 \tag{8}$$

and use the following linear interpolation to smooth the amplitude.

$$\tilde{A}_l(n) = \hat{A}_l^k + \frac{(\hat{A}_l^{k+1} - \hat{A}_l^k)}{S} n \tag{9}$$

where in Equation (8), $\tilde{\theta}_l(t)$ is a smoothed phase for the l th sine wave of current frame. t=0 corresponds to the current frame, and t=T corresponds to the next frame. T is the frame length of synthetic speech. a, b, c and d are all coefficients to be solved. In Equation (9), $\tilde{A}_l(n)$ is the smoothed amplitude for the l th sine wave of current frame. $\hat{A}_l^k$ and $\hat{A}_l^{k+1}$ are the estimated amplitudes for the l th sine wave of current frame and next frame respectively.

Solving Equation (8) using the method in reference [4], we can obtain:

$$\tilde{\theta}_l(t) = \hat{\theta}_l^k + \hat{\omega}_l^k t + c(M^*)t^2 + d(M^*)t^3 \tag{10}$$

where $\hat{\theta}_l^k$ and $\hat{\omega}_l^k$ are respectively the l th estimated component of phase and frequency for the current frame. $c(M^*)$ and $d(M^*)$ are the solved value of coefficients c and d after optimization. For those matched frequency components, we can use the above formulas to smooth the phase and amplitude directly. For those unmatched components, we should first have some initial value for the estimated amplitude and phase components, then we can use the above formulas to smooth them. That is to say, if there is a frequency component that ends in the next frame, then the amplitude for the next frame is set to be zero, and phase can be represented by the following formula:

$$\hat{\theta}_l^{k+1} = \hat{\theta}_l^k + \hat{\omega}_l^k S \tag{11}$$

If there is a frequency component that starts in the next frame, then amplitude of current frame is set to be zero, and phase can be expressed to be the following formula:

$$\hat{\theta}_l^k = \hat{\theta}_l^{k+1} - \hat{\omega}_l^{k+1} S \tag{12}$$

After smoothed, the synthetic speech for the current frame can be represented using the following formula:

$$\tilde{s}(n) = \sum_{l=1}^{L^k} \tilde{A}_l(n) \cos[\tilde{\theta}_l(n)] \tag{13}$$

where L^k is the number of auditory spectrum lines in the current frame.

4. EXPERIMENTS

In this paper, speech is pre-processed in the same way as in reference [1], but the frame length is different and is set to be 25ms and without any overlapping between the frames. Using same quantization and coding scheme as in reference [1], the average coding bit-rate is 2.6kbps if there 5 spectrum lines at average for a frame of speech, and the highest bit-rate is 4.16kbps if the maximal spectrum lines is limited to be 8. The principle of speech synthesis with parametric matching and smoothing is presented in Figure 5.

Figure 6 and Figure 7 present a segment of original clean speech and its corresponding synthetic speech respectively. Figure 8 and Figure 9 present a segment of original noisy speech and its corresponding synthetic speech where Figure 8 corresponds to Figure 6.

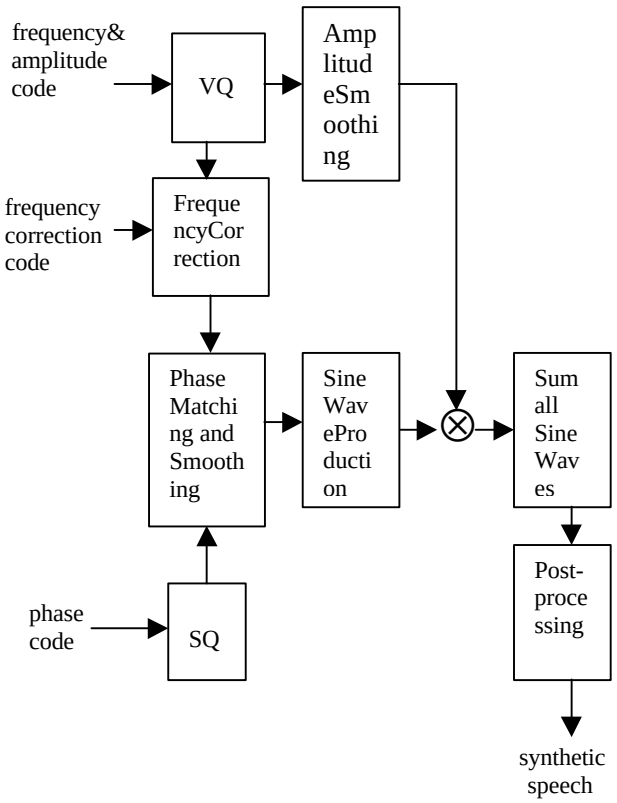

Figure 5. Principle of speech synthesis with parametric matching and smoothing

Figure 6. A segment of original clean speech

Figure 7. Synthetic speech corresponding to Figure 6

Figure 8. A segment of original noisy speech corresponding to Figure 6 with SNR=15dB

Figure 9. Synthetic speech corresponding to Figure 8

From Figure 7 and Figure 9, we can find that two segments of synthetic speech look quite similar from the appearance. It shows that the proposed algorithm is quite robust in the noisy background with SNR higher than 15dB, the quality of synthetic speech is no worse than that presented in reference [1].

5. CONCLUSION

This paper applies a parametric matching and smoothing method to the sinusoidal representation and auditory model-based speech analysis/synthesis system. Without the change of synthetic speech quality, the bit-rate for speech coding is reduced from 3.25kbps to 2.6kbps. In the noisy background (white noise) with SNR higher than 15dB, the synthetic speech quality is almost unchanged. Experiments show that the proposed algorithm has better robustness and adaptation than the conventional ones.

Acknowledgements

We acknowledge the collaboration of Yuan Jingxian of Shanghai University and Yim Chiho of Hong Kong University of Science and Technology.

6. REFERENCES

[1] Wan Wanggen, Au C.L.Oscar, ”A novel approach of low bit-rate speech coding based on sinusoidal representation and auditory model”, EUROSPEECH’99 (to appear).

[2] O. Ghitza, “Auditory nerve representation criteria for speech analysis/ synthesis”, IEEE Trans. Acoust., Speech and Signal Processing, Vol.35, No.6, p.736-740, 1987.

[3] Wan Wanggen, Yu Xiaoqing, ”A second-order difference cochlear model”, Acta Electronica Sinica, Vol.23, No.7, p.6-9, 1995(in Chinese).

[4] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation”, IEEE Trans. Acoust., Speech and Signal Processing, Vol.34, p.744-754, August 1986.