A Sinusoidal Harmonic Vocoder at 1.2 kbps Using Auditory Perceptual Characteristics

Minoru KOHATA

Chiba Institute of Technology, Narashino, 275-0016, Japan, kohata@net.it-chiba.ac.jp

ABSTRACT

In this paper, a very low bit speech coder at 1.2 kbps is newly proposed. Like the LPC vocoder, it requires few types of information (power, pitch, and spectral information), but its quality is far superior. In the proposed vocoder, the synthesized speech quality is improved based on auditory perceptual characteristics. The synthesis method is one of harmonic coding, using sinusoids whose frequencies are multiples of the fundamental frequency, where the amplitudes of the sinusoids are adaptively modulated using Gammatone filters as a perceptual weighting filter. The sinusoids' phases are also adjusted so as to maximize the perceptual quality. In order to reduce the total bit rate to 1.2 kbps, a new segment coder for spectral information (LSP coefficients) using DP matching is also proposed. The quality of the synthesized speech is considerably improved compared with that of the simple LPC vocoder, according to MOS and preference tests.

1. INTRODUCTION

Some of the recent low bit speech coders use an architecture based on the classical vocoder, though the quality of the synthesized speech is considerably improved compared with that of a simple vocoder. A 2.4 kbps or lower coder can use only a small amount of information to represent speech, so it is quite difficult to preserve good quality for CELP[1] type coders. The recent low bit coders might be classified into WI[2] (waveform interpolation), MELP[3] (mixed excitation linear predictive coding), or harmonic coding. In these coders, though the synthesized speech waveform does not exactly follow the input, the subjective quality is preserved through some perceptual redundancy reductions. Among those coders, the harmonic coder is considered to be the easiest with which to implement human auditory characteristics.

In this paper, a 1.2 kbps coder based on a "perceptual harmonic coder" is newly proposed. This coder uses sinusoids whose amplitude and phases are modulated to improve subjective quality of the synthesized speech. Then various phase and amplitude modulation methods were tested and compared through subjective listening tests. In order to reduce the total bit rate to 1.2 kbps, a new low bit spectral coding method was also proposed. Finally, the proposed 1.2 kbps coder was simulated and the synthesized speech quality evaluated; it outperformed an LPC vocoder by 0.7 in MOS tests, and by 31% in preference tests.

2. PERCEPTUAL HARMONIC SYNTHESIS

2.1. CSW method

The speech synthesis method of the proposed coder is similar to that of a harmonic coder[4], which sums up sinusoids whose frequencies are multiples of F0, and synthesizes speech signals. The original harmonic coder, as proposed by Tribolett, controls each frequency of the sinusoids precisely, and phase also is controlled. The synthesis method employed here is a simplified version of the harmonic coder. The difference from the original is that no additional information other than power, F0, and LSP is required. In this method, only the continuity of these parameters is ensured and the remaining information for the perceptual modulation is explicitly given at a receiver. We call this the CSW (continuous sinusoidal waveform) method. In the CSW method, synthesized speech is represented as Eq. (1),

$$s(t)=\sum_{i=1}^{N} a_i(t)\sin\{i\omega_{p(t)}+\phi_i(t)\} \qquad (1)$$

where $a_i(t)$, $\phi_i(t)$, and $\omega_{p(t)}$ represent the amplitude, the phase, and the angular pitch frequency, respectively. N is the number of sinusoids to be added, which is determined by the Nyquist frequency and the pitch frequency. If $a_i(t)$ were set to the spectral envelope obtained by LPC analysis, and all $\phi_i(t)$ were set to zero or random phase, this coder would be an LPC vocoder. Our aim in the following is to perceptually control these parameters adequately to improve the "buzzy" quality of the LPC vocoder.

2.2. Perceptual phase modulation

Figure 1 shows a block diagram of the proposed 1.2 kbps coder. Here, how to insert the phases of the sinusoids at the receiver is described. Firstly Eq. (1) is modified as Eq. (2),

$$s(t)=\sum_{i=1}^{N} a_i(t)\sin\{i\omega_{p(t)}+\phi_i'(t)+\varphi_i(T)\} \qquad (2)$$

where $\varphi_i(T)$ denotes the i-th sinusoid's phase at the end of the previous frame; this ensures phase continuity between adjacent frames. And $\phi_i'(t)$ represents the phase variation in the present frame. This phase information affects the speech waveform in a pitch period. It is often said that human auditory perception is not so sensitive to this phase

information, but the speech quality is definitely enhanced if the phase information is decided carefully.

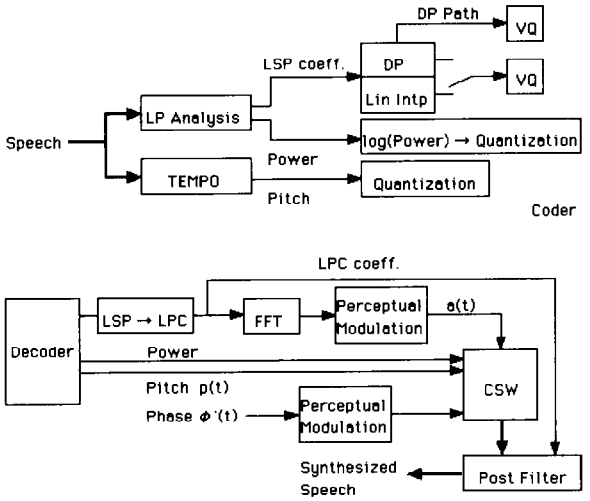

Fig. 1 Speech coder using CSW method.

Various methods to supply phase information at the receiver without increasing the number of bits were tried. As a result, the speech quality is relatively good in methods (2) and (3) as follows. Method (1) is the control.

(1) Set all $\phi_i'(t)$ to zero.

(2) Substitute the minimum phases for $\phi_i'(t)$.

(3) Substitute the harmonics' phases of the Rosenberg pulse for $\phi_i'(t)$, which is obtained by sampling the FFT phase spectrum of the Rosenberg pulse.

Figure 2 shows the harmonic phases of the Rosenberg pulse in method (3). In method (1), the perceptual effect of the phase is not considered and it's quality is equivalent to that of the LPC vocoder. These three methods were compared through a preference test with thirty-six sentences uttered by six different speakers. The results are shown in Table 1, with method (3) giving the best quality. The speech synthesized by method (3) was felt to be fairly natural and buzzy-less compared with that of method (1).

2.3. Perceptual amplitude modulation

As shown in Fig. 1, the amplitudes of the sinusoids are calculated from LSP coefficients by using FFT at the receiver. If these amplitudes were applied directly to the CSW method, the synthesized speech might be felt to be as buzzy as that of an LPC vocoder. This buzzy quality is caused by the complete harmonic structure of the spectrum. In MBE[5] or MELP, the excitation signal is composed of a mixture of impulse and noise, and the ratio of the mixture is determined adaptively by each sub-band, in order to avoid buzzy-ness and to improve perceptual quality. But these coders require additional bits to control the mixture. Here, how to improve the perceptual quality without additional bits is described. Unlike in MELP, the proposed method does not use mixing. Rather, the amplitudes of the sinusoids of the CSW method are adaptively modulated considering the perceptual quality.

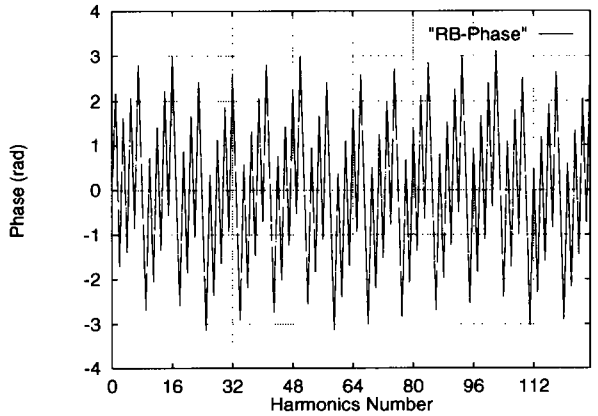

Fig. 2 Harmonic phases of Rosenberg pulse.

Phase control method	Preference score (%)
Set all to zero	36.0
Minimum phase	43.7
Rosenberg	70.0

Table 1 Subjective quality vs. phase control method.

Experiments with various amplitude modulation methods were conducted, and according to informal listening tests, the following two methods gave better quality.

(1) Modulation with A-level weighting function: A-level weighting function is measured as the ratio of a perceptual sound intensity to a physical sound intensity, related with frequency. This function is shown in Fig. 3, and denoted by $A(f)$. The amplitudes $a_i(t)$ are linearly decreased to zero in the present frame, if $a_i(t)$ satisfies Eq. (3).

$$a_i(t) < Th \cdot \max_{j=1,N}\{a_j(t)\} / A(f_i) \tag{3}$$

Th is a constant to determine the threshold, and f_i is the harmonic frequency corresponding to the i-th harmonic sinusoid.

(2) Modulation with Gammatone filters[6]: The former method with A-level weighting function modulates the amplitudes independently from the spectral structure of the input speech since the threshold is determined by the maximum value of $a_i(t)$. Then the Gammatone filter is introduced to modulate the amplitudes depending on the spectral structure. The Gammatone filter is one of the filters which simulate the auditory perceptual characteristics. The characteristic of the Gammatone filter is given by Eq. (4),

$$GT_i(f) = A(f_i)\left[1 + j\frac{f - f_i}{b}\right]^{-n} \tag{4}$$

where n = 4, b = 1.019 ERB on the ERB[7] frequency scale. This characteristic is shown in Fig. 4.

The Gammatone filters are used to make a function which substitutes for $A(f)$ in Eq. (3). This function is calculated as Eq. (5).

$$A_G(f) = \sum_{i=1}^{N} \int_0^{4kHz} H(f) \cdot GT_i(f) df \quad (5)$$

where $H(f)$ is the LPC amplitude spectrum. This function reflects the perceptual auditory sensitivity, which depends on a temporal spectral structure. The amplitude of the sinusoids are modulated the same way as in the former method,.

By an informal listening test, the synthesized speech sounds quite natural and buzzy-less in both modulation methods. In order to compare these two methods, a preference test was conducted with the same speech samples in the preference test for phase modulation. In both methods, the Rosenberg harmonic phase is applied. The results are shown in Table 2, and the method with Gammatone filters is clearly superior to the others.

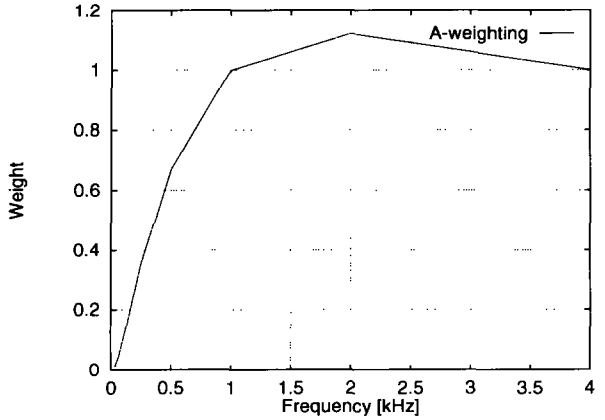

Fig. 3 A-weighting characteristics.

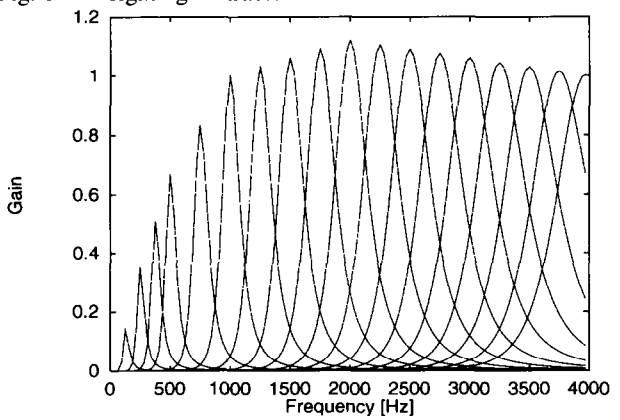

Fig. 4 Frequency characteristics of Gammatone filters.

Threshold control method	Preference score (%)
None	44.6
A-weighting	56.7
Gammatone filter	73.0

Table 2 Subjective quality by threshold control method.

3. QUANTIZATION

3.1. LSP quantization

The proposed coder requires the quantization of power, pitch frequency (F0), and LSP coefficients. In order to reduce the total bit rate to 1.2 kbps, the LSP coefficients must be quantized efficiently. Here, a new quantization method for LSP coefficients, named LIN-DP, which can reduce the bit rate for LSP to 450 bps is described.

(1) Linear interpolation: Firstly, LSP coefficients of order 10 of the input speech are calculated each 8 ms to build a segment which has constant length. Then linear interpolation is carried out between each top frame of the segment. The sampling frequency of the input speech is 8 kHz.

(2) DP matching: Secondly, DP matching is carried out, where the input pattern is the linear interpolated segment and the template pattern is the original segment. The DP path is restricted by Eq. (6).

$$g(i,j) = \min \begin{bmatrix} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + d(i,j) \\ g(i-1,j-2) + d(i,j) \end{bmatrix} \quad (6)$$

Where $g(i-1,*)$ represents the sum of the distance up to the i-1th frame, and $d(i,j)$ is the distance at the i-th frame. For quantization, the frequently used DP paths are recorded in a codebook, and the quantization is carried out in the same way as for VQ.

DP matching is applied only if Eq. (7) is satisfied, that is, if enough gain from the DP matching is obtained.

$$\frac{D_L}{H} > D_p \quad (7)$$

H is a positive constant larger than 1, D_L and D_p are the distortion in the linear interpolation and in the DP matching, respectively. This selective use of DP matching reduces the total bits needed for the DP path quantization.

As a result, the LIN-DP quantizes the top frame of the segment with split-VQ (10 bit + 10 bit), and outputs one bit of the DP switch bit. If the DP switch is on, then three more bits are added to quantize the DP path. Fig. 5 shows the cepstral distortion in LIN-DP and in linear interpolation against bit rate. The bit rate is varied by changing the segment length from 48 ms to 96 ms. In all cases, LIN-DP quantizes LSP coefficients with lower distortion than linear interpolation. We have adopted the conditions corresponding to the point of 450 bps, 2.26 dB in Fig. 5.

3.2. Quantization of remaining parameters

Table 3 shows the bit allocation for the proposed coder. As described in the above subsection, for LSP quantization, 24 or 21 bits are allocated depending on whether or not DP is used. The pitch frequency is obtained by the TEMPO[8] algorithm proposed by Kawahara, and quantized linearly with 7 bits each 16 ms. The LPC residual power is used for the power in the synthesis, whose logarithm is scalarly quantized with 5 bits each 16 ms.

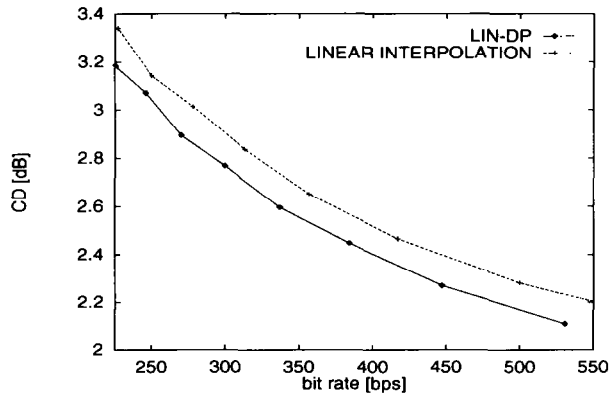

Fig. 5 Cepstral distortion by LIN-DP.

Parameters	Rates	
LSP's	10 + 10 bit/segment (Split VQ)	450 bps
DP switch	1 bit/segment	
DP path	3 bit/segment	
Pitch	7 bit/frame	750 bps
Power	5 bit/frame	
Total		1.2 kbps

Table 3 Bit allocations for the proposed 1.2 kbps coder.

4. SUBJECTIVE TESTS

Finally, simple subjective listening tests were carried out with six sentences uttered by six different speakers. These tests consisted of preference tests and MOS tests with five levels (1-5) of scoring.

The synthesized speech to be evaluated is listed in Table 4. Speech #4 in Table 4 was synthesized with decreased power for unvoiced frames. This improves the quality by decreasing the explosive noise perceived in a rising consonant period. The results are shown in Table 5. In the preference test, the quality of speech #2 is clearly preferred over speech #1, which was not perceptually modulated. After quantization, the quality of speech #3 is slightly degraded, but it remains better than that of #1. Most of this degradation is assumed to be caused by the quantization of LSP. Speech #4 shows a little improvement compared with #3, and this demonstrates that the power control in a consonant frame is efficient.

Speech	Conditions of synthesis
#1	CSW only
#2	Rosenberg + Gammatone
#3	Quantized #2 (1.2 kbps)
#4	#3 + power control in UV frame

Table 4 Conditions of speech synthesis.

Speech #	# 1	# 2	# 3	# 4
Preference score (%)	42.9	62.8	46.8	48.7
MOS	2.34	3.18	2.55	2.80

Table 5 Results of the preference and MOS tests.

In the MOS tests, results similar to those in the preference tests are obtained.

In order to confirm the effects of the perceptual modulation, the proposed method was compared with an ordinary LPC vocoder operating at the same rate. The conditions for quantization are the same as for those of the proposed method. The results are shown in Table 6. The proposed coder could clearly outperform the ordinary LPC vocoder.

	LPC vocoder	Proposed 1.2 kbps coder
Preference score (%)	34.7	65.3
MOS	2.1	2.8

Table 6 Subjective qualities of the proposed method and LPC vocoder.

5. CONCLUSIONS

In this paper, a new harmonic coder operating at 1.2 kbps using auditory perceptual characteristics is proposed. Perceptual phase modulation using Rosenberg pulse's harmonic phase, and perceptual amplitude modulation using Gammatone filters greatly improved the quality of the simple harmonic vocoder. The quality is not yet as good as that of a 2.4 kbps coder such as MELP, but we were able show the possibility of coding speech at such a very low bit rate using perceptual modulations.

6. REFERENCES

1. M.R.Schroeder and B.S.Atal, "Code-Exited Linear Prediction (CELP):High Quality Speech at Very Low Bit Rates," *Proc.ICASSP*, pp.937-940, 1985.

2. W.B.Kleijin and J.Haagen, "A Speech Coder Based on Decomposition of Characteristic Waveforms," *Proc.ICASSP*, pp.508-511,1995.

3. Alan V.MacCree and Thomas P.Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," IEEE Trans. on Speech and Audio Processing, Vol.3, No.4, pp.242-250, 1995.

4. Jorge S. Marques, Luis B.Almedia and Jose M.Tribolet, "Harmonic Coding at 4.8 kb/s," *Proc ICASSP*, pp.17-20, 1990.

5. Tian Wang, Kun Tang and Chongxi Feng, "A High Quality MBE-LPC-FE Speech Coding at 2.4 kbps and 1.2 kbps," *Proc. ICASSP*, pp.208-211, 1996.

6. de Boer E. and de Jongh H.R., "On cochlear ending: Potentialities and limitations of the reverse correlation technique," J.Acoust.Soc.Am., vol.63,1, pp.115-135,1978.

7. Patterson R.D. et al., J.Acoust. Soc.Am., vol.98, pp.1890-1894, 1995.

8. http://www.hip.atr.co.jp/~kawahara/STRAIGHT.html