Speech Quality of a Sinusoidal Model Depending on the Number of Sinusoids

Jeong Wook Seo* · Ki Hong Kim · Jong Won Seok*** · Keun Sung Bae***

ABSTRACT

The STC(Sinusoidal Transform Coding) is a vocoding technique that uses a sinusoidal speech model to obtain high quality speech at low data rate. It models and synthesizes the speech signal with fundamental frequency and its harmonic elements in frequency domain. To reduce the data rate, it is necessary to represent the sinusoidal amplitudes and phases with as small number of peaks as possible while maintaining the speech quality. As a basic research to develop a low-rate speech coding algorithm using the sinusoidal model, in this paper, we investigate the speech quality depending on the number of sinusoids. By varying the number of spectral peaks from 5 to 40 speech signals are reconstructed, and then their qualities are evaluated using spectral envelope distortion measure and MOS(Mean Opinion Score). Two approaches are used to obtain the spectral peaks: one is a conventional STFT (Short-Time Fourier Transform), and the other is a multiresolutional analysis method.

Keywords : speech coding, STC, speech quality, sinusoidal model

I. INTRODUCTION

It has been reported that an analysis and synthesis system based on a sinusoidal representation of speech leads to synthetic speech that is perceptually indistinguishable from the original speech[1]. The STC(Sinusoidal Transform Coding) is a frequency domain speech compression technique, in which short-time segments of the speech signal are represented by linear combination of sinusoids with time-varying amplitudes, phases, and frequencies. The STC, as a vocoding scheme, is known to produce reconstructed speech of good quality at data rates below 10 kbps[2]. For

* School of Electronic and Electrical Eng., Kyungpook National Univ., Taegu, Korea
** LG Electronics Co.
*** ETRI, Broadcasting Technology Department, Radio & Broadcasting Technology Lab.
**** This work was done by University Research Program supported by Ministry of Information & Communication in Korea.

this class of vocoders, speech is analyzed first by segmenting the signal using a finite duration analysis window and spectral peaks that correspond to sinusoidal components are selected. In the sinusoidal model, the input speech signal at the kth segment is represented as the sum of a finite number of sinusoidal components, and given by

$$s^k(n) = \sum_{m=1}^{M} A_m^k \cos(w_m^k n + \phi_m^k) \quad (1)$$

where M is the number of sinusoidal components in the speech bandwidth, and A_m^k, ϕ_m^k, and w_m^k denote the time-varying amplitude, phase and frequency of each underlying sine wave. As the speech is analyzed from frame to frame, different sets of the above parameters are obtained[3].

Since as many as 50 or more spectral peaks can exist for a typical low-pitched speaker in the 4 kHz speech bandwidth, it is not possible to code all of the parameters directly in the speech coding system. Therefore it is necessary to decrease the number of sinusoidal components in the synthesis procedure while maintaining the speech quality. In this paper, by varying the number of sine waves from 5 to 40, speech signals are reconstructed and then their qualities are evaluated using the cepstral distance as a spectral envelope distortion measure and MOS(Mean Opinion Score) as a subjective test.

The rest of this paper is organized as follows. In the next section, the sinusoidal model of a speech signal is explained and a brief description of multiresolutional analysis is given. In section 3, two speech quality measures, cepstral distance and MOS, are explained. Some experimental results are presented with our findings and discussion in section 4, and conclusion is given in section 5.

II. SINUSOIDAL MODEL OF SPEECH SIGNAL

In the speech production model, the speech waveform is assumed to be the output of passing a glottal excitation waveform through a linear system representing the characteristics of vocal tract. The excitation function is usually represented as a periodic pulse train during voiced speech, where the spacing between consecutive pulses corresponds to the pitch of the speaker while it is noise-like signal during unvoiced speech.

In the sinusoidal model of a speech signal, the binary voiced/unvoiced excitation is represented by the sum of sine waves. The motivation for this sine wave

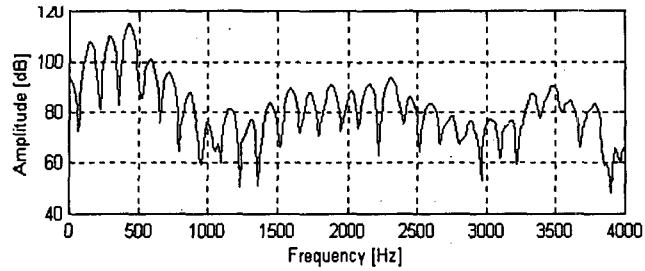
representation is that nearly periodic voiced excitation can be represented by harmonic components that correspond to sine waves. The sinusoidal model represents speech by a linear combination of sine waves with time-varying amplitudes, frequencies, and phases[4]. This model is used to represent finite duration segments of a speech signal. Let $s^k(n)$ denote a windowed segment of a speech signal as given in Eq. (1). If the speech is assumed to be periodic, the sine wave parameters correspond to the harmonic samples of the STFT. In this case, Eq. (1) can be rewritten by

$$s^k(n) = \sum_{l=1}^{L} A_m^k \cos(nlw_o^k + \phi_l^k) \tag{2}$$

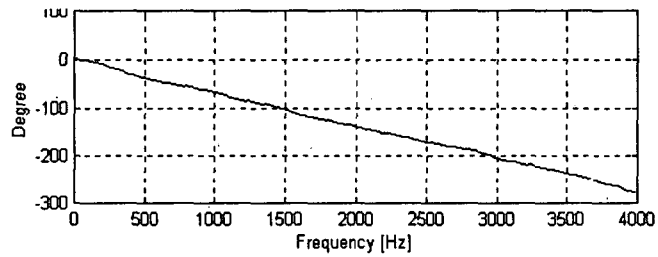
where the sine wave frequencies are multiples of the fundamental frequency w_o and the corresponding amplitude and phase are given by the harmonic samples of STFT. If the STFT of $s(n)$ is given by

$$S(w) = \sum_{n=-N/2}^{N/2} s(n) e^{-jnw} \tag{3}$$

then the Fourier analysis gives the amplitude estimates as $A_l = |S(lw_o)|$, and the phase estimates as $\phi_l = \arg[S(lw_o)]$. The magnitude of the STFT will have peaks at multiples of w_o. Figures 1 and 2 show the spectral characteristics of speech signal in the voiced and unvoiced regions. Figure 1 shows the speech waveform in the voiced region, and we can see spectral peaks corresponding to the harmonics of fundamental frequency, especially in the formant area. But in figure 2 of the unvoiced speech signal, we can see many spectral peaks that have irregular intervals in the frequency domain. Therefore, more sine waves are needed to model the speech in the unvoiced/transition region.

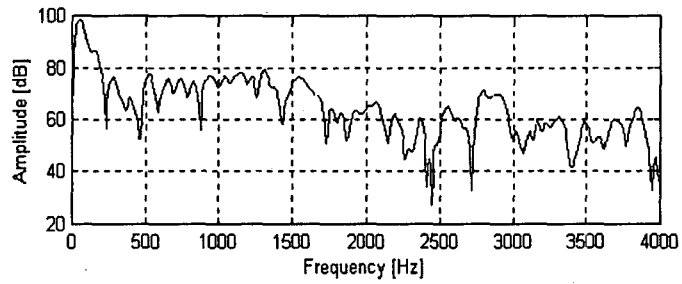


(a) Amplitude

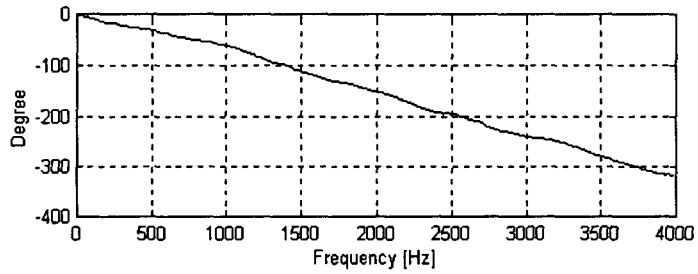


(b) Phase

Figure 1. Spectral characteristics in the voiced region



(a) Amplitude



(b) Phase

Figure 2. Spectral characteristics in the unvoiced region

Figure 3 shows the block diagram of the STC analysis/synthesis system. In analysis system, first, the input waveform is windowed and discrete Fourier transform(DFT) is performed. Peak-picking is then applied to the magnitude spectrum of the DFT in order to obtain a list of frequencies and corresponding amplitudes at those frequencies. The total number of spectral peaks per frame are adaptively limited based on the expected average pitch period of the utterance. If the amplitudes, frequencies, and phases that are estimated for kth frame are denoted by A_l^k, w_l^k, and ϕ_l^k, then the synthetic speech for that frame can be computed using the following equation.

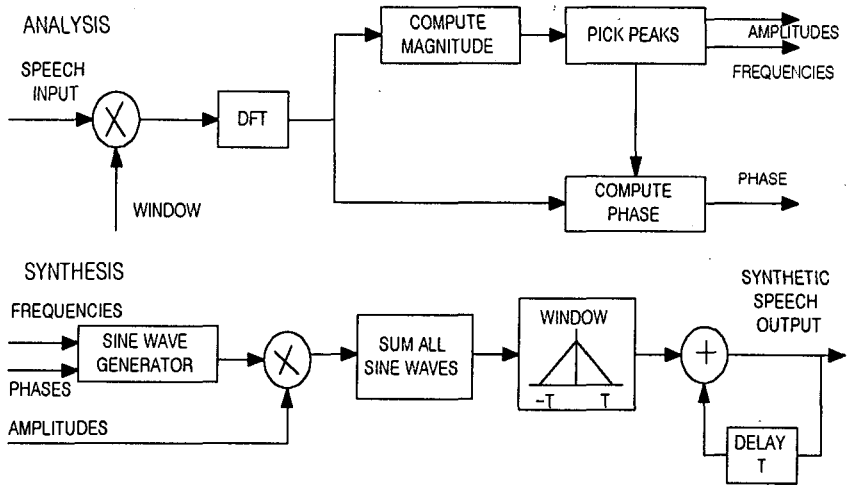


Figure 3. Block diagram for analysis/synthesis of sinusoidal modeling system

$$\hat{s}^k(n) = \sum_{l=1}^{L} A_l^k \cos(nw_l^k + \phi_l^k) \tag{4}$$

Since the sine wave parameter will be time-varying, discontinuities at the frame boundaries will be introduced. To overcome this problem, overlap-and-add interpolation with a window function is generally used[5]. This method is simple to implement and satisfactory for the speech coding application provided that the synthesis frame is made short enough to satisfy the stationarity assumption of sine wave parameters. In this case, the synthetic speech waveform is obtained from Eq. (5).

$$\hat{s}(n) = w_s(n)\ \hat{s}^{k-1}(n) + w_s(n-T)\ \hat{s}^{k}(n-T) \quad (5)$$

where $\hat{s}^{k-1}(n)$ is reconstructed spech segment of($k-1$)th frame, and $w_s(n)$ is the overlap-and-add synthesis window function that satisfies Eq. (6). Triangular, Hanning and trapezoidal windows have typically been used for this overlap- and-add process.

$$w_s(n) + w_s(n-T) = 1 \quad (6)$$

We use two different approaches for obtaining spectral peaks from the segment of speech signal. One is finding spectral peaks directly from the DFT of the speech signal. The other is from the DFT of the wavelet transformed speech signal. Wavelet transform enables us kind of multiresolutional spectral analysis. It is known that a dyadic discrete wavelet transform results in the octave band filter bank analysis, i.e., nonuniform subband filtering and its spectral characteristics are very similar to the perceptual characteristics of human being[6]. In other words, lower frequency sinusoidal components are calculated over a greater length of time with better frequency resolution and higher frequency sinusoids are calculated with poor frequency resolution but high time resolution.

There is inherently some degree of aliasing in the subband signals that could be cancelled by the synthetic filter bank with perfect reconstruction condition[7,8]. If the sinusoidal modeling is applied in each scale prior to synthesis, however, aliasing cancellation in the reconstruction is not guaranteed. This aliasing cancellation problem can be removed, if a nondecimated wavelet transform is used to split the input signal into required bands. The output of nondecimated wavelet transform satisfy the perfect reconstruction constraint such that where q is the number of scales to split the input signal. Eq. (7) means that there is no aliasing introduced in the subband signals and each subband signal has the same length. In this scheme, spectral peaks are obtained in each scale and multiresolution is achieved by using windows of different duration in each scale. Figure 4 shows the structure for analysis and synthesis with multiresolutional sinusoidal modeling.

$$\sum_{q} x_q(n) = x(n) \quad (7)$$

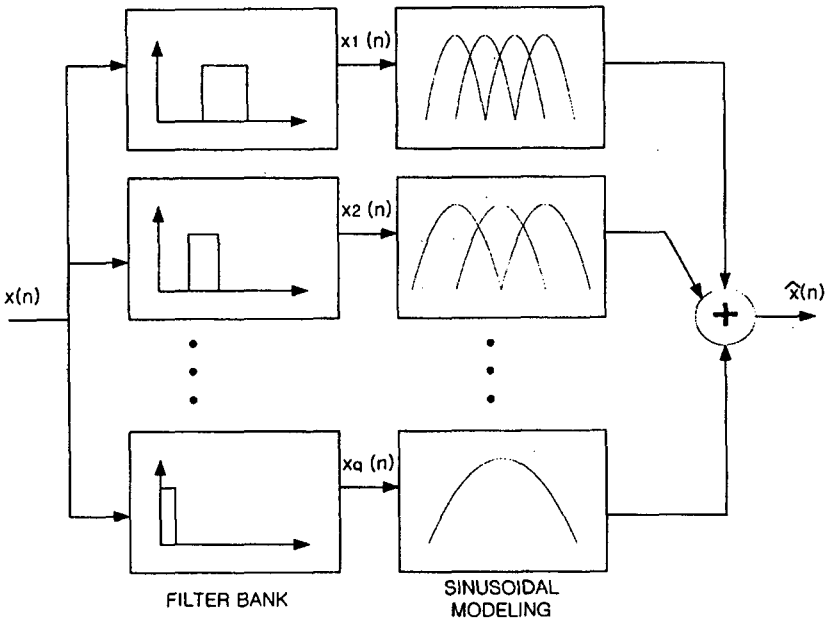


Figure 4. Structure for analysis/synthesis with multiresolutional sinusoidal modeling

III. OBJECTIVE AND SUBJECTIVE SPEECH QUALITY MEASURE

In this paper, we use MOS as a subjective speech quality measure. MOS test[9], which is one of the successive methods in psychometric measurements, is employed to assess directly the naturalness of speech quality. In the case of telecommunications systems, five grades of speech quality are often distinguished: excellent(E), good(G), fair(F), poor(P), and unsatisfactory(U). Their weighted mean values are calculated using 5 to 1, respectively, for each grade.

For the objective speech quality test, we use cepstral distance(CD)[9], which is one of spectral envelope distortion measure. It has been reported that, in the frequency domain measures, spectral envelope distortion measures have better correlation with subjective assessment values than the spectral distortion measures that consider a whole speech spectrum. Thus the CD measure is thought to be appropriate for MOS estimation. The CD based on the LPC-cepstrum is given in Eq. (8),

$$CD^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K}[C_x(i) - C_y(i)]^2 \qquad (8)$$

where K is cepstrum order, N is frame number, and $C_x(i)$ and $C_y(i)$ is cepstrum coefficient of original and synthesized signals, respectively.

Ⅳ. EXPERIMENTAL RESULTS

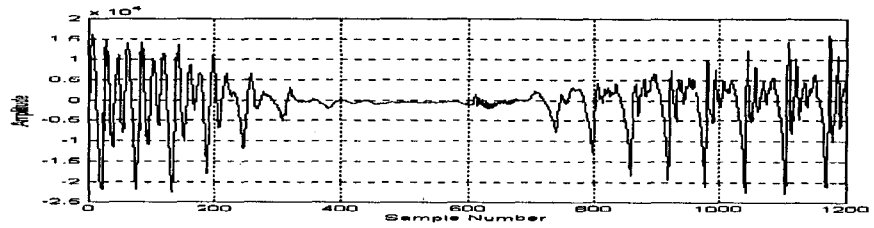
The utterances of two Korean sentences and one English sentence were sampled at 8 kHz with 16 bits quantization per sample. Analysis frame size was set to 20 msec, and 1024-point FFT per frame was used for spectral analysis. In spectral analysis, we used Hamming window, and Triangular window was used in the synthesis procedure.

For the nondecimated wavelet transform, we used the Daubechies 10-tap filter. And, input speech which has 40 ms frame size is split into three subbands using nondecimated wavelet transform. Ranging from the lowest to highest band, the subband sinusoidal modeling used window size of 40, 20, 10 ms, respectively. The number of sinusoids for each subband were determined experimentally from each scale in the wavelet domain.

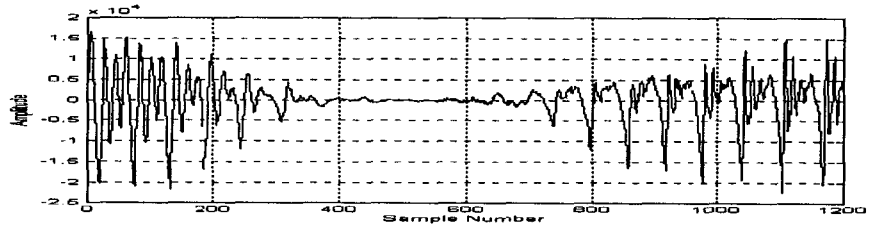
Figure 5 shows original signal and synthesized signal according to harmonic number, i.e., the number of sine waves used for synthesis, and figure 6 shows the spectral characteristics of them. In figure 5, it is shown that as the harmonic number for reconstructing signal decreases, the difference between original and synthesized signal waveform increases rapidly. When 20 harmonics or more sine waves were used for reconstructing the signal, the synthesized signals maintained similar waveform and speech quality as the original one.

Figure 6 shows the spectral characteristics of synthesized signal according to harmonic number. It can be seen that as the harmonic number decreases, formant information lost in the high frequency region increases more. Similar to the results of figure 5, when we used 20 harmonics or more for reconstructing signal, the synthesized signal contains most of the formant components and phase information.

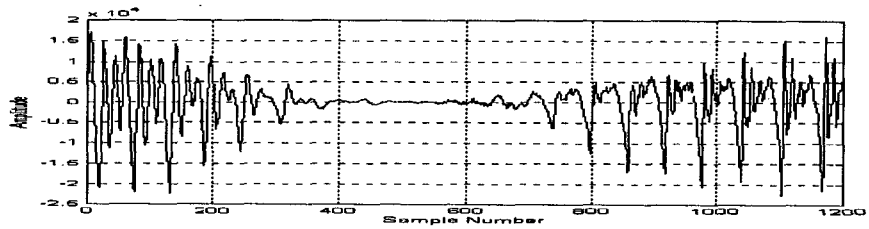
Table 1 shows the results of objective speech quality assessment, i.e., the cepstral distance of reconstructed speech signals. The results of the subjective speech quality test, MOS, are shown in table 2. From these tables, it is shown that around 20 of sinusoids are necessary to achieve approximately toll quality of speech signals. On the contrary to our expectations, the results of multiresolutional analysis with wavelet transform was slightly poorer than the conventional one.



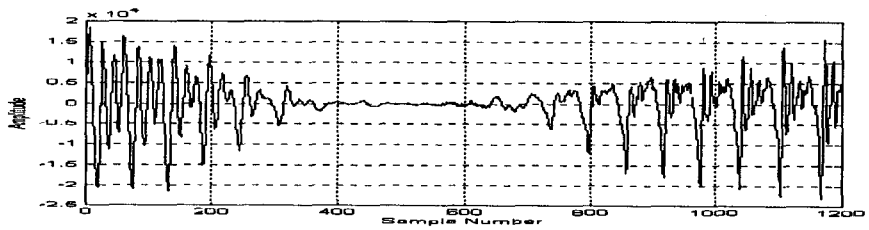
(a) Original signal



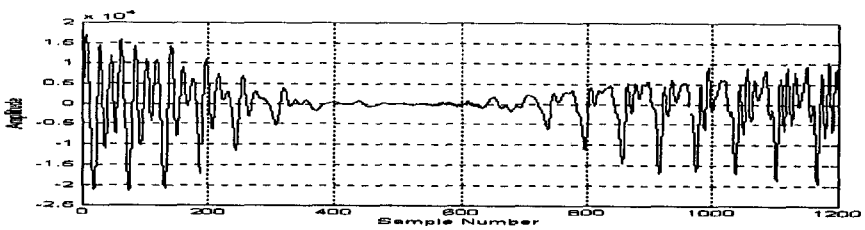
(b) Synthesized signal using 40 harmonics



(c) Synthesized signal using 20 harmonics

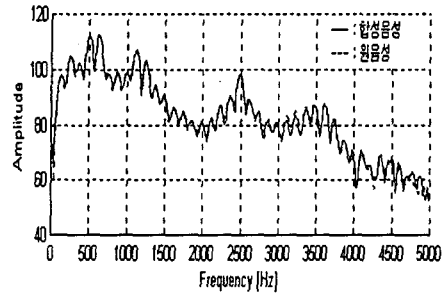


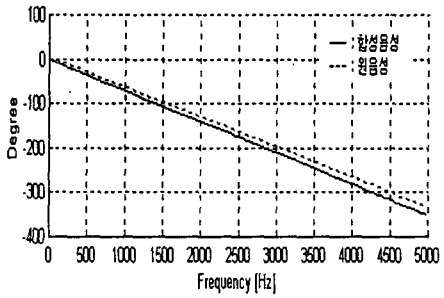
(d) Synthesized signal using 10 harmonics



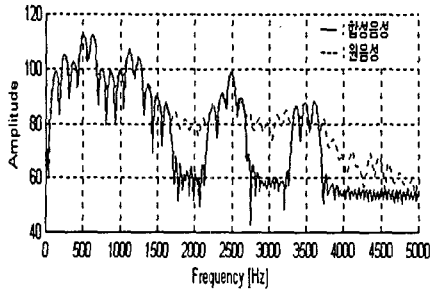
(e) Synthesized signal using 5 harmonics

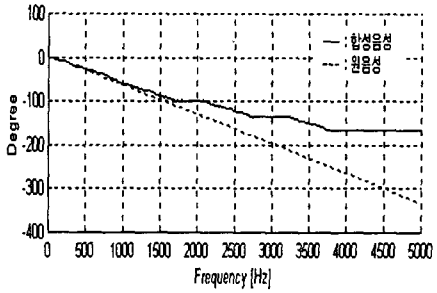
Figure 5. Comparison of original and synthesized signals according to harmonic number



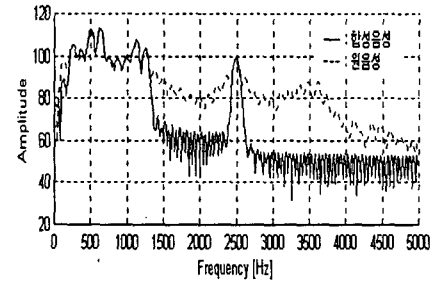


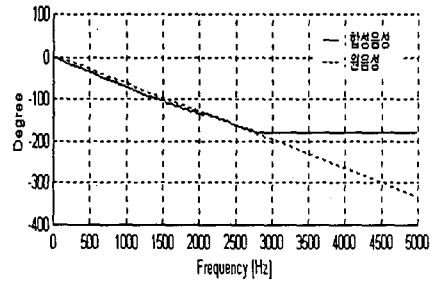
(a) 40 harmonics are used



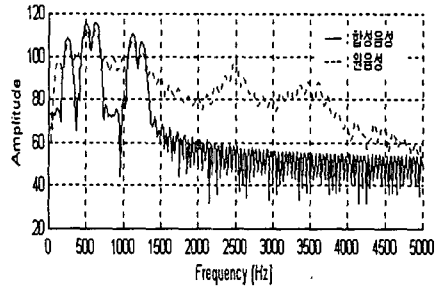


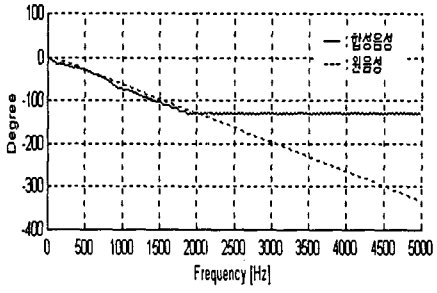
(b) 20 harmonics are used





(c) 10 harmonics are used





(d) 5 harmonics are used

Figure 6. Comparison of spectral characteristics according to harmonic number

Table 1. Results of cepstral distance

number of sinusoid	sentence 1		sentence 2		sentence 3	
	DFT	wavelet	DFT	wavelet	DFT	wavelet
5	1.6	1.5	1	1.1	1	1.4
10	2.2	2.4	1.9	1.6	2.2	2.4
15	3.2	3.0	2.9	2.5	3.4	2.8
20	3.7	3.7	3.7	3.6	3.8	3.6
25	4.3	3.9	3.9	3.8	4.4	3.8
30	4.3	4.2	4.4	4.1	4.7	4.2
40	4.7	4.4	4.4	4.6	4.8	4.4

Table 2. Results of MOS test

number of sinusoid	sentence 1		sentence 2		sentence 3	
	DFT	wavelet	DFT	wavelet	DFT	wavelet
5	0.0599	0.1878	0.1034	0.1604	0.1204	0.1640
10	0.0513	0.1499	0.0912	0.1455	0.0836	0.1354
15	0.0501	0.0622	0.0883	0.1357	0.0776	0.0981
20	0.0502	0.0616	0.0877	0.1349	0.0755	0.0960
25	0.0501	0.0616	0.0876	0.1349	0.0740	0.0954
30	0.0501	0.0616	0.0877	0.1349	0.0732	0.0955
40	0.0500	0.0507	0.0874	0.1324	0.0727	0.0859

V. CONCLUSION

The STC is a vocoding technique that uses a sinusoidal speech model to obtain high quality speech at low data rate. As a basic research to develop a low-rate speech coding algorithm using the sinusoidal model, we investigated the speech quality depending on the number of sinusoids. By varying the number of spectral peaks from 5 to 40, speech signals were reconstructed, and then their qualities were evaluated using the cepstral distance and MOS. Spectral peaks were obtained using two approaches, short-time Fourier transform, and multiresolutional analysis method.

Experimental results demonstrated that as the harmonic number for reconstructing signal decreases the difference between original and synthesized signal waveform increases rapidly. When 20 harmonics or more sine waves were used for reconstructing the signal, the synthesized signals maintained similar waveform and speech quality

compared to the original one. Similarly, results of cepstral distance and MOS test also showed that about 20 of sinusoids were necessary to achieve toll quality of synthetic speech. On the contrary to our expectations, the results of multiresolutional analysis with wavelet transform was slightly poorer than the conventional one. Further studies on the multiresolutional analysis and harmonics and phase modeling are being undertaken.

REFERENCES

[1] R. J. McAulay, T. F. Quatieri. 1986. "Phase Modeling and Its Application to Sinusoidal Transform coding", *ICASSP'86*, 1713-1715.

[2] R. J. McAulay, T. F. Quatieri. 1995. *Speech Coding and Synthesis*, W. B. Kleijn, and K. K. Paliwal Eds., Elsevier.

[3] R. J. McAulay and T. F. Quatieri. 1986. "Speech Analysis/ Synthesis Based on a Sinusoidal Representation", *IEEE Trans. on ASSP*, Vol.34, No.4, 744-754.

[4] T. F. Quatieri and R, J, McAulay 1986. "Speech Transforms Based on Sinusoidal Representation." *IEEE Trans. on ASSP*, Vol.34, No.6, 1449-1464.

[5] E. B. George and M. J. T. Smith. 1997. "Speech Analysis using and Analysis/ Overlap-Add Sinusoidal Model." *IEEE Trans. on ASSP*, Vol.5, No.5, 389-406.

[6] I. Daubechies. 1992. *Ten Lectures on Wavelets*, SIAM.

[7] D. Anderson. 1996. "Speech Analysis and Coding Using a Multiresolutional Sinusoidal Transform." *ICASSP'96.*

[8] O. Rioul, M. Vetterrli. 1991. "Wavelet and Signal Processing," *IEEE Signal Processing Magazine*, 14-38.

[9] S. Furui, M. M. Sondhi. 1991. *Advances in Speech Signal Processing*, Marcel Dekker, Inc.,

Received : Jan. 10, 2000.
Accepted : Feb. 22, 2000.

▲ Jeong Wook Seo
School of Electronic and Electrical Engineering,
Kyungpook National University, Taegu, Korea
Tel.: +82-53-940-8627, Fax: +82-53-950-5505
E-mail: jwseo@palgong.knu.ac.kr

▲ Ki Hong Kim
LG Electronics, Display Division
E-mail: hong0612@lge.co.kr

▲ Jong Won Seok
ETRI, Broadcasting Technology Department,
Radio & broadcasting Technology Lab.
Tel.: +82-42-860-1306 Fax.: +82-042-860-6465
E-mail: jwseok@etri.re.kr

▲ Keun Sung Bae
School of Electronic and Electrical Engineering,
Kyungpook National University, Taegu, Korea
Tel.: +82-53-950-5527, Fax.: +82-53-950-5505
E-mail: ksbae@ee.knu.ac.kr