A MULTI−RESOLUTION SINUSOIDAL MODEL USING ADAPTIVE ANALYSIS FRAME

Ki-Hong Kim, In-Ho Hwang

National Security Research Institute
P.O.BOX 1, Yuseong, Daejeon, Korea
phone: +82 42 860 1425, fax: +82 42 860 5656, email: hong0612@etri.re.kr
web: www.nsri.re.kr

ABSTRACT

The sinusoidal model has been applied to speech and audio signal coding, analysis/synthesis, modification, and various other fields. However, one drawback of this model is that the analysis frame size is generally fixed in analysing the speech and audio signal. As a result, optimal spectral resolution cannot be guaranteed to each sinusoidal component. In this paper, in order to overcome this drawback and to estimate the sinusoidal components more accurately, we propose a multi-resolution sinusoidal model using an adaptive analysis frame size. In the proposed scheme, the analysis frame size is varied based on the coarse and refined pitch characteristics. Experimental results have shown that the proposed model can achieve better performance than that of classical and multi-resolution sinusoidal models using a filter bank.

1. INTRODUCTION

The sinusoidal model has been applied to a broad range of speech/audio signal coding, analysis/synthesis, and various other fields since it was first proposed in the 1980s[1-7]. Since it is easy to develop a proper method to manipulate its components {amplitudes, frequencies, phases}, the sinusoidal model has proven useful for speech and audio time- and frequency-scale modifications, fundamental frequency modifications, speech enhancement, and co-channel separation[6-8]. STC(Sinusoidal transform coding), which is based on a sinusoidal model of a speech signal, is known to achieve a better quality of synthetic speech than that of CELP(Code Excited Linear Prediction) at data rates around 2.4 to 4.8Kbps[3-5].

The sinusoidal model represents a speech signal as a linear combination of sinusoids with time-varying amplitudes, frequencies, and phases. That is, in voiced regions, the speech signal is represented as the sum of a finite number of corresponding sinusoidal components at the fundamental frequency and its harmonics. In unvoiced regions, it is represented as the sum of numbers of corresponding sinusoidal components at the peaks in the spectral domain[1-7].

In the classical sinusoidal model, the analysis window size is generally fixed in analysing the speech and audio signal[1-8]. However, since each sinusoidal component has different frequencies, an analysis window with fixed size cannot guarantee optimal spectral resolution to each sinusoidal component and thus it results in lowered coding gain and reconstruction artifacts such as pre-echo[9-11].

In this paper, in order to overcome these drawbacks in the classical sinusoidal model and to estimate the sinusoidal components more accurately, we propose a multi-resolution sinusoidal model using an adaptive analysis frame size in the sinusoidal model. This approach applies a variable-size window to the speech signal. That is, the analysis frame size is varied based on the coarse and refined pitch characteristics using a specific pitch estimation[12,13].

The remainder of this paper is organized as follows. In the next section, a detailed description of the classical sinusoidal model is given. In section 3, an existing multi-resolution analysis using a filter bank and the proposed multi-resolution analysis using an adaptive analysis frame size are illustrated. Some experimental results are presented in section 4, and concluding remarks are provided in section 5.

2. CLASSICAL SINUSOIDAL MODEL

The speech signal is assumed to be the output of a vocal cord excitation signal passed through a linear system representing the characteristics of the vocal tract. The excitation signal is usually represented as a periodic pulse train during voiced speech, and is represented as a noise-like signal during unvoiced speech. In the sinusoidal model, the binary voiced/unvoiced excitation model can be replaced by a sum of sine waves. That is, the excitation signal is represented as the sum of a finite number of corresponding sinusoidal components at the fundamental frequency and its harmonics during voiced regions, and is represented as numbers of corresponding sinusoidal components at peaks in the spectral domain during unvoiced regions[1-7].

In the sinusoidal model, the input speech signal is represented as

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l) \qquad (1)$$

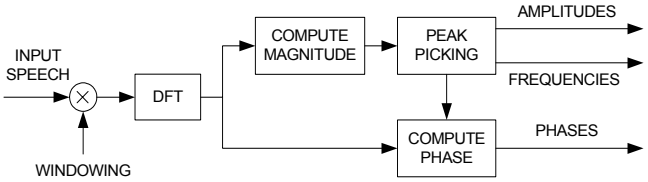
where L is the number of sinusoidal components, and A_l, ω_l and ϕ_l represent the time-varying amplitude, frequency, and phase of each underlying sine wave, respectively. If the amplitudes, frequencies, and phases that are estimated for the k th frame are denoted by A_l^k, ω_l^k and ϕ_l^k respectively,

the synthetic speech $\tilde{s}^k(n)$ for that frame can be computed using

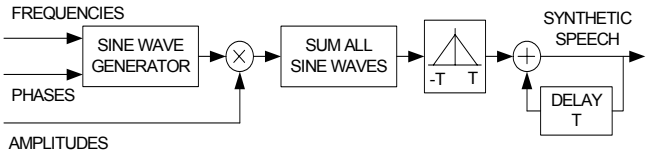
$$\tilde{s}^k(n) = \sum_{l=1}^{L^k} A_l^k \cos(\omega_l^k n + \phi_l^k) \quad (2)$$

Since the sinusoidal components are time-varying, discontinuities at the frame boundaries are introduced. To overcome this problem, the overlap-add method is used.

Figure 1 shows a block diagram of the classical sinusoidal model analysis/synthesis system. In the analysis system, amplitudes, frequencies, and phases are found. Then, in the synthesis system, synthetic speech is reconstructed with these sinusoidal components.



(a) Analysis system



(b) Synthesis system

Figure 1: Block diagram of classical sinusoidal model.

3. MULTI-RESOLUTION SINUSOIDAL MODEL

In the classical sinusoidal model, an analysis window size generally has fixed size[1-8]. This is not very effective in terms of achieving higher coding gain. Another problem caused by the fixed-size window is difficulty in modelling noise-like components and time-localized transient regions, and these result in pre-echo distortion in the synthetic speech signal[9-11]. To overcome these problems, some multi-resolution models have been suggested[9-11]. Such multi-resolution models can be achieved by two methods : filter bank method[9], or adaptive windowing method [10,11], wherein the analysis frame size is varied based on the signal characteristics.

3.1 Multi-Resolution Model Using Filter Bank

The sinusoidal model has been reinterpreted using the multi-resolution approach, which utilizes the concept of a filter bank to overcome the problem in the classical model. In the filter bank method, first the input speech signal is decomposed into subband signals using a filter bank. Lower frequency components are then calculated over a greater length of time and a finer frequency resolution. Higher frequency components are estimated with poor frequency resolution but high time resolution.

Figure 2 shows a block diagram of the multi-resolution sinusoidal model using a filter bank. In this scheme, the input speech signal is decomposed into subband signals using a filter bank. These signals are then independently analysed and synthesized with the classical sinusoidal model. But since this model independently analyses and synthesizes subband signals, it has a computational disadvantage.

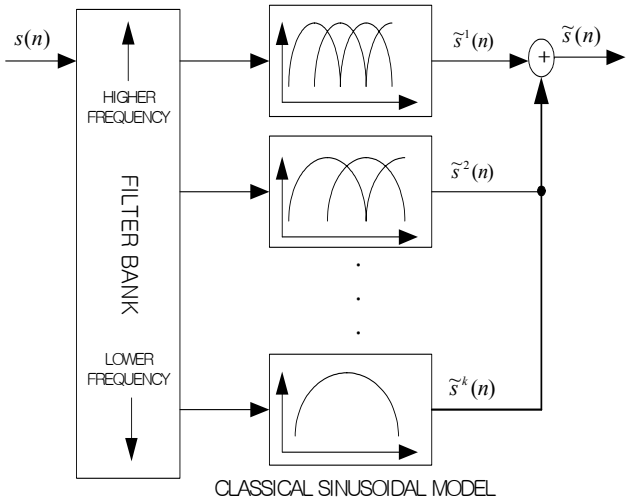


Figure 2: Block diagram of multi-resolution model using filter bank.

3.2 Proposed Multi-Resolution Model Using Adaptive Analysis Frame

To overcome the problems of the classical model and to estimate sinusoidal components more accurately, we propose a multi-resolution sinusoidal model using an adaptive analysis frame size method in the classical model.

In the proposed multi-resolution sinusoidal model, coarse pitch analysis is first performed to estimate the coarse pitch[12]. In the first step, three maxima of the correlation given by eq. (3) are found in three ranges, and these are normalized by eq. (4).

$$R(k_i) = \sum_{n=101}^{200} s(n)s(n-k_i), \quad \begin{matrix} i=1,\ k_1 = 80,...,143 \\ i=2,\ k_2 = 40,...,79 \\ i=3,\ k_3 = 20,...,39 \end{matrix} \quad (3)$$

$$R'(t_i) = \frac{R(t_i)}{\sqrt{\sum_n s^2(n-t_i)}},\ i = 1,2,3 \quad (4)$$

The winner among the three normalized correlations $(R', t_i),\ i = 1,2,3$ is selected by favoring the delays in the lower ranges. A specific refined pitch analysis[13] is then performed to estimate the refined pitch, and this is represented as

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \overline{A(k\omega_0)} \left\{ \max_{\omega_l \in L(k,\omega_0)} [A_l D(\omega_l - k\omega_0)] - \frac{1}{2} \overline{A}(k\omega_0) \right\} \quad (5)$$

$$D(\omega - k\omega_0) = \frac{\sin[2\pi(\frac{\omega - k\omega_0}{\omega_0})]}{2\pi(\frac{\omega - k\omega_0}{k\omega_0})} \quad (6)$$

$$L(k\omega_0) = \{\omega : k\omega_0 - \frac{\omega_0}{2} \le \omega \le k\omega_0 + \frac{\omega_0}{2}\} \qquad (7)$$

where A_l, $\overline{A}(\omega)$, and ω_l represent the all peaks amplitudes, spectral envelope, and its frequencies, respectively.

The final pitch P is obtained using the coarse pitch P_{coarse} of the current frame estimated coarse pitch analysis and the refined pitch $P_{refined}$ of the previous frame, and then the adaptive windowing size W_{len} can be determined using

$$P = \frac{P_{coarse} + P_{refined}}{2} \qquad (8)$$

$$W_{len} = 2.5 \times P,\ 140 \le W_{len} \le 200 \qquad (9)$$

Figure 3 shows a block diagram of the proposed multi-resolution sinusoidal model, and Figure 4 shows the adaptive windowing structure of this model. In the analysis system, first a voiced/unvoiced decision[13] is performed using a coarse pitch. In voiced regions, sinusoidal components are found using a final pitch adaptive windowing size, and in unvoiced regions, these are found using a windowing size used in the voiced region of the previous analysis frame. Then, in the synthesis system, a synthetic speech signal is reconstructed with these sinusoidal components.

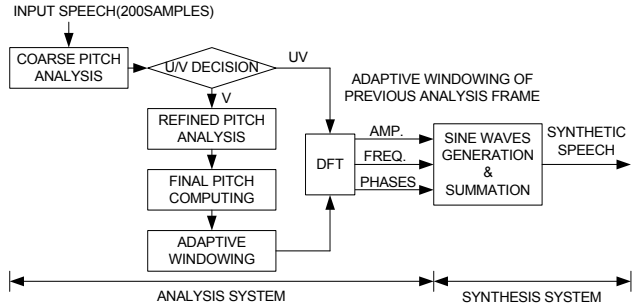


Figure 3: Block diagram of proposed multi-resolution sinusoidal model.

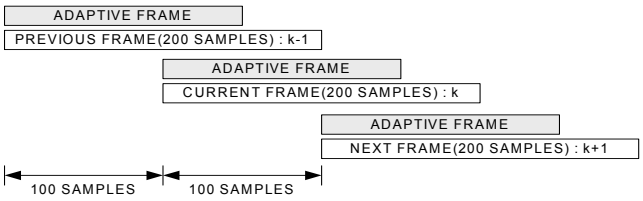
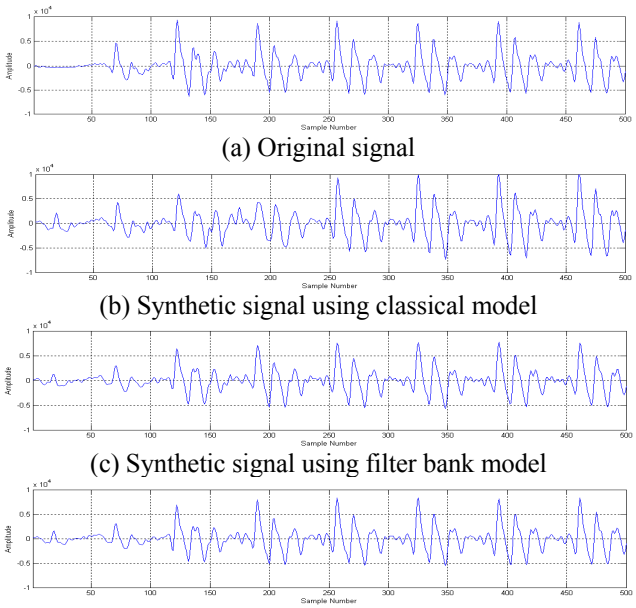


Figure 4: Adaptive window structure.

4. EXPERIMENTAL RESULTS

In this paper, we compared and analysed the waveform and spectrum of original and synthetic speech depending on the number of sinusoids. The performance of the proposed model has been evaluated in terms of the cepstrum distance (CD) measure[3] and ITU Rec. P.862, PESQ test[14] based on objective speech quality measures. Ten speech signals were sampled at 8KHz with 16bits quantization per sample. In the classical model, the analysis frame size is set to 25msec, and 1024 points FFT are used. In the analysis, we used a hamming window, and in the synthesis, a triangular window was used. For the multi-resolution model using a filter bank, we used a quadrature mirror filter(QMF). We decomposed an input speech signal of 50ms frame size into 3 subbands. Ranging from the lowest to highest band, the subband sinusoidal model uses window sizes of 50, 25, and 12.5ms.

Figure 5 shows one division of original and synthetic speech signals using the classical and filter bank approaches as well as the proposed model, and Figure 6 shows a spectrum of the original and synthetic speech signals.



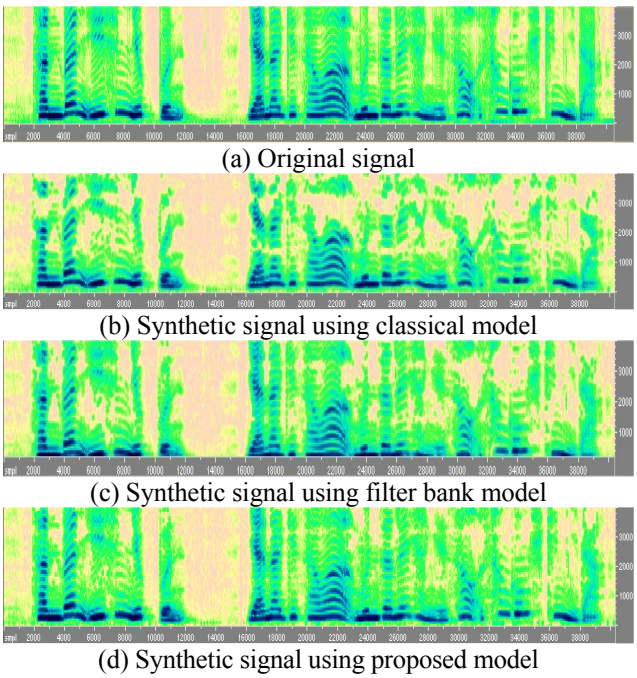
(a) Original signal

(b) Synthetic signal using classical model

(c) Synthetic signal using filter bank model

(d) Synthetic signal using proposed model

Figure 5: Waveform between original and synthetic speech.



(a) Original signal

(b) Synthetic signal using classical model

(c) Synthetic signal using filter bank model

(d) Synthetic signal using proposed model

Figure 6: Spectrum between original and synthetic speech.

In these figures, we see that the synthetic speech signal waveform using the proposed model corresponds more closely to the original signal than those using the classical and filter bank model. It is also demonstrated that the spectrum of the synthetic speech signal from the proposed multi-resolution model more accurately approximates that of the original speech signal.

Table 1 shows results of objective speech quality assessments, CDs of synthetic speech signal. From this table, we can see that CDs decrease as sinusoids number increases, and that CDs using the proposed multi-resolution model are smaller than those using the classical and filter bank method.

Table 1. Cepstrum distance of the synthetic speech signal.

Sentence	Model	Sinusoids number			
		15	20	30	40
1	Classical	0.7741	0.5534	0.3339	0.1370
	Filter bank	0.5915	0.4733	0.2472	0.0505
	Proposed	0.5693	0.4589	0.1545	0.0501
2	Classical	0.8874	0.4579	0.2304	0.1163
	Filter bank	0.4767	0.4007	0.1879	0.0843
	Proposed	0.3605	0.2726	0.1503	0.0854
3	Classical	0.8046	0.3848	0.1613	0.0586
	Filter bank	0.4686	0.3547	0.1096	0.0759
	Proposed	0.4156	0.3187	0.0974	0.0454
4	Classical	0.8670	0.3350	0.0939	0.0711
	Filter bank	0.2494	0.1879	0.0453	0.0397
	Proposed	0.1881	0.1297	0.0412	0.0312
5	Classical	0.6339	0.2234	0.1160	0.0847
	Filter bank	0.1701	0.1159	0.0557	0.0465
	Proposed	0.0985	0.0589	0.0578	0.0472
6	Classical	0.5908	0.2684	0.1582	0.0826
	Filter bank	0.2183	0.1678	0.0629	0.0452
	Proposed	0.1327	0.0454	0.0450	0.0417
7	Classical	0.8322	0.3423	0.2220	0.0745
	Filter bank	0.2431	0.1753	0.0876	0.0377
	Proposed	0.1461	0.0710	0.0675	0.0238
8	Classical	0.5687	0.3863	0.1658	0.0653
	Filter bank	0.3273	0.2377	0.0528	0.0381
	Proposed	0.2872	0.2075	0.0387	0.0351
9	Classical	0.8331	0.4109	0.2244	0.0953
	Filter bank	0.2899	0.2101	0.0930	0.0439
	Proposed	0.2661	0.1783	0.0671	0.0412
10	Classical	0.8344	0.3500	0.1816	0.0920
	Filter bank	0.3536	0.2657	0.0654	0.0580
	Proposed	0.3137	0.2195	0.0370	0.0365

The results of the PESQ test are shown in Table 2. A simulation for comparison of synthetic speech quality shows that the proposed model improves MOS over 0.079~0.21 compared with the classical model according to the sinusoids number.

Table 2. PESQ test of the synthetic speech signal.

Model	Sinusoids number			
	15	20	30	40
Classical	3.330	3.422	3.592	3.628
Filter bank	3.357	3.502	3.724	3.764
Proposed	3.406	3.597	3.777	3.808

5. CONCLUSIONS

In order to enhance performance of the estimation process in the classical sinusoidal model and multi-resolution sinusoidal model using a filter bank, we proposed and implemented a multi-resolution sinusoidal model using an adaptive analysis frame. In the proposed scheme, the analysis frame size is varied based on the coarse and refined pitch characteristics.

Experimental results showed that the proposed sinusoidal model achieved better performance in terms of waveform, spectrum characteristics, and synthetic speech quality.

The proposed multi-resolution sinusoidal model provides natural speech signal decomposition, high quality output, and a combination of modification flexibility. It can be applied to many applications including speech and audio signal processing, low rate coding, and modification.

REFERENCES

[1] R.J.McAulay & T.F.Quatieri, "Speech Analysis/Synthesis Based on Sinusoidal Representation", *IEEE Trans. on ASSP*, vol. 34, pp. 744-754, 1986.

[2] T.F.Quatieri & R.J.McAulay, "Speech Transform Based on a Sinusoidal Representation", *IEEE Trans. on ASSP*, pp. 1449-1464, 1986.

[3] S.Furui & M.M.Sondhi, *Advances in Speech Signal Processing*, Dekker Inc., NY, 1992.

[4] K.N.Hamdy, M.Ali & A.Tewfik, "High Quality Audio Coding of Audio Signal with a Combined Harmonics and Wavelet Representation", *IEEE ICASSP*, pp. 1045-1048, 1996.

[5] W.B.Kleijn & K.K.Paliwal, *Speech Coding and Synthesis*, Elsevier, 1995.

[6] E.B.George & M.J.T.Smith, "Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/ Overlap-Add Sinusoidal Model", *IEEE Trans. on ASSP*, vol. 5, pp. 389-406, 1997.

[7] E.B.George & M.J.T.Smith, "Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones", *Journal of Audio Eng. Soc.*, vol. 40, pp. 497-516, 1992.

[8] T. F.Quatieri & R.G.Daisewicz, "An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech", *IEEE Trans. on ASSP*, vol. 38, pp. 56-69, 1990.

[9] D.V.Anderson, "Speech Analysis and Coding Using a Multi-Resolution Sinusoidal Transform", *IEEE ICASSP*, pp. 1037-1040, 1996.

[10] S.N.Levine, T.S.Verma, & J.O.Smith III, "Alias-free, Multiresolution Sinusoidal Modeling for Polyphonic, Wideband Audio", *IEEE ASSP*, pp. 19-22, 1997.

[11] M.Goodwin, "Multiresolution Sinusoidal Modeling Using Adaptive Segmentation", *IEEE ICASSP*, pp. 1525-1528, 1998.

[12] R.Salami, C.Laflamme, J.P.Adoul, & D.Massaloux, "A Toll Quality 8 Kb/s Speech Codec for the Personal Communications System(PCS)", *IEEE Trans. on Vehicular Technology*, vol. 43, pp. 808-816, 1994.

[13] R.J.McAulay & T.F.Quatieri, "Pitch Estimation and Voicing Detection Based on A Sinusoidal Speech Model", *IEEE ICASSP*, pp. 249-252, 1990.

[14] ITU-T Rec. P.862, Perceptual Evaluation of Speech Quality(PESQ) an Objective Assessment of Narrowband Telephone Networks and Speech Codes, 2002.