Proc. Biennial Symp. Commun. (Kingston, Ont.), pp. 29-32, May 1992

Speech Coding Using an Enhanced Sinusoidal Model at Low Bit-Rate

Qian Yasheng[1,2] Liu Jia, Feng Chongxi[1] Peter Kabal[2]

[1]Department of Electronics Engineering
Tsinghua University
Beijing, China
100084

[2]INRS-Telecommunications
Université du Québec
3 Place du Commerce
Verdun, Quebec
Canada H3E 1H6

Abstract

An enhanced sinusoidal model, which employs the time-varying amplitudes of three components to track the fast dynamical variations during the transition speech segments, and exploits the redundancies between the near-neighborhood components to reduce the number of sinusoidal components to a maximum of 20 with high synthesized quality is presented. Many components can be determined by linear prediction of the dominant and fundamental components, thereby reducing the number of the parameters required to be transmitted and the corresponding bit rate. This approach improves the synthesized quality of the unvoiced and transition speech segments.

An optimal algorithm for extracting dominant frequencies by formats and pitches is compared with a DFT method. The effects on the synthesis quality of the number of the time-varying amplitudes and the different base functions are compared.

Two vector quantization codebooks with group classifications are developed to reduce the storage and computation load for a 4.8 kbits/s coder. Objective measurements give a cepstrum distance of 2.62 dB for several phonetically balanced sentences. Informal listening tests have shown that the proposed speech coder with an enhanced sinusoidal model can obtain good quality speech at 4.8 kbits/s.

1. Introduction

A speech coding system with high quality, at bit rates from 2.4 to 4.8 kbits/s, will be required for the next generation of the advanced mobile radio and satellite communications systems. The sinusoidal analysis/synthesis system proposed by McAulay [1], has been shown to produce synthetic speech of very high quality. However, a direct application of this method requires a large number of the parameters in the basic sinusoidal representation to achieve high quality speech. Several speech coding system, based on harmonic coding [2] and zero-phase version [3] of the sinusoidal model, have been recently introduced to reduce the required number of the parameters of the model. These modifications result, to some extent, in the degradation of the speech quality, particularly in unvoiced and transition segments.

An enhanced sinusoidal model, which employs the time-varying amplitudes of three components to track the fast dynamical variations and exploits the redundancies between the adjacent components to reduce the number of sinusoidal parameters with high quality is presented in this paper. Many components can be reconstructed by the linear prediction of the dominant and fundamental components. This approach improves the speech quality, particularly for unvoiced and transition segments, at low bit rates.

The proposed enhanced sinusoidal model is described in Section 2. An optimal algorithm for extracting dominant frequencies is discussed in Section 3. The effects on the speech quality of the number of the time-varying amplitudes and the different base functions are compared in Section 4. In the last section a parameter quantization strategy is described for a 4.8 kbits/s coder. Objective measurements give a cepstrum distance of 2.62 dB for several phonetically balanced sentences. Informal listening tests have shown that the proposed speech coding system can obtain high quality at 4.8 kbits/s.

2. The Enhanced Sinusoidal Model

A basic sinusoidal representation of speech signal $S(t)$ for the frames of length T could be approximately expressed by a sum of sine waves with arbitrary amplitudes, frequencies and phases. For n-th frame,

$$\hat{S}(t) = \sum_{k=1}^{N_n} A_{k,n}(t)\cos(\omega_{k,n}t + \theta_{k,n}(t)) \qquad (1)$$

Where $A_{k,n}(t)$ — the amplitude of the k-th sinusoidal component

$\theta_{k,n}(t)$ — the phase of the k-th sinusoid

N_n — the number of sinusoidal components

Since speech signals are not truly periodical, the number of the sinusoids required for unvoiced and transition segments could be as many as 80. The objective of the enhanced sinusoidal model is to reduce the maximum number of sinusoids from 80 to a smaller number such as 20, while maintaining high quality.

The enhanced sinusoidal model can be represented by (for frame n)

$$\begin{aligned}\hat{S}(t) &= \sum_{k=1}^{N} A_{k,n}(t) \sum_{i=-D}^{D} a_{ki}\cos(\omega_{k,n}t + i\omega_{p,n}t + \theta_{k,n}(t)) \\ &= \sum_{k=1}^{N} A_{k,n}(t)\cos(\omega_{k,n}t + \theta_{k,n}(t)) \\ &\quad + \sum_{k=1}^{N}\sum_{i=-D,i\neq 0}^{D} A_{k,n}(t)a_{ki}\cos(\omega_{k,n}t + i\omega_{p,n}t + \theta_{k,n}(t))\end{aligned} \qquad (2)$$

where $\hat{S}(t)$ — the reconstructed speech signal

$\omega_{k,n}$ — dominant frequencies

$A_{k,n}(t)$ — time-varying amplitudes

$\theta_{k,n}(t)$ — time-varying phases

$\omega_{p,n}$ — a pitch frequency or estimate of unvoiced segment

Part of this research was supported by a grant from the Canadian Institute for Telecommunications Research under the NCE program of the Government of Canada.

$a_{k,1}$	$a_{k,2}$	$a_{k,3}$	$a_{k,4}$	$a_{k,5}$
0.2	0.1	0.025	0.01	0.01

Table 1 Spectral Prediction Coefficients

$\{ a_{k,i} \}$— a set of spectral prediction coefficients
D — the number of the prediction components
N — the number of the dominant sinusoids

The difference between the basic and the enhanced sinusoidal model is the second term in the expression (2). Only the parameters of the dominant sinusoids need be transmitted. The other sinusoids are generated by spectral regeneration using linear spectral prediction, as shown in Fig. 1. The sinusoids between the dominant components, that is those with frequencies $\omega_{k,n}-m\omega_{p,n}$, $\omega_{k,n}+(-m+1)\omega_{p,n}$, ..., $\omega_{k,n}+(m-1)\omega_{p,n}$, $\omega_{k,n}+m\omega_{p,n}$, are reconstructed by linear prediction of the corresponding $2D$ neighboring dominant sinusoidal components at frequencies $(\omega_{l-D,n}, \omega_{l-D+1,n}, \ldots, \omega_{l+D-1,n}, \omega_{l+D,n})$.

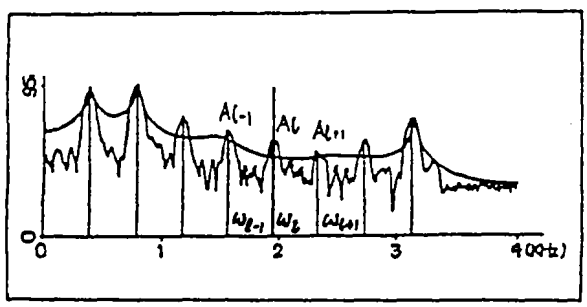

Fig. 1 Spectral Regeneration

The estimates of the m regenerated components between the dominant frequencies are determined by the expression,

$$\hat{A}_l = \sum_{i=-D, i\neq 0}^{D} a_{k,i} A_{l+i} \tag{3}$$

where $\{ a_{k,i} \}$ are the spectral prediction coefficients.

The set of the spectral prediction coefficients are determined with the mean-square-error criterion. Let E denote the MSE for the difference between the real and estimated amplitudes.

$$E = \sum_{l=1}^{N} [A_l - \hat{A}_l]^2 = \sum_{l=1}^{N} [A_l - \sum_{i=-D, i\neq 0} a_{k,i} A_{l+i}]^2 \tag{4}$$

The minimization of the MSE of (4) leads to a set of linear equations (5) with a Toeplitz matrix.

$$\sum_{i=1}^{D} a_{k,i} \Phi(i,m) = \alpha(m), \qquad m = \{-D, D\} \tag{5}$$

In this paper we use 5 fixed spectral prediction coefficients. The equations have to be solved from the statistics, based on a large speech data base. In addition, experiments were conducted to fine tune the parameters based on subjective quality. The resulting coefficients are shown in Table 1.

3. Estimation of Time-varying Amplitudes in a frame

The time-varying amplitudes for the sinusoidal model within the frames can track the fast dynamical variations during the transitions. Those amplitudes are specified by the linear combinations of some known functions of time t,

$$A_k(t) = \sum_{i=0}^{Q_k} c_{i,k} u_i(t) \qquad k = 1, 2, ..., N$$
$$u_i(t) = t^i \tag{6}$$

where the subscript k refers to the k-th sinusoid and $c_{i,k}$ are the parameters to be estimated.

The amplitudes can be well approximated with this polynomial, as the number of terms Q_k increases. The amplitudes are determined by analysis-by-synthesis procedures to fit speech envelopes. Substituting (6) into the sinusoidal representation (2), the synthesized signal can be given as

$$\hat{S}(t) = \sum_{i=1}^{N} g_i f(i,t) \tag{7}$$

where the g_i are the parameters, which include the both phases and amplitudes through simple trigonometric manipulations.

$$f(i,t) = \begin{cases} t^i \sum_{i=-D}^{i=D} a_{k,i} \sin(\omega_k t + i\omega_k t) \\ t^i \sum_{i=-D}^{i=D} a_{k,i} \cos(\omega_k t + i\omega_k t) \end{cases} \tag{8}$$

where $1 \leq k \leq N$, $1 \leq i \leq Q_k$.

The estimation of the set of g_i coefficients is performed by the minimization of the mean-square-error E between the original and synthesized signals,

$$E = \sum_{n=0}^{L-1} [S(t) - \hat{S}(t)]^2 \tag{9}$$

It leads to a set of linear equations with positive definite matrix, which could be efficiently solved by the Cholesky decomposition method.

Six different parameters for the time-varying amplitudes have been tested and compared with segmental SNR (dB) and subjective listening. The experimental results are listed in Table 2. Using six sinusoids with one linear and two quadratic varying amplitudes is much better than ten sinusoids with constant amplitudes. The improvement in terms of SNR improvement is 5.85 dB. The larger the number of sinusoids with time-varying amplitudes, the better the objective and perceptual quality. However, the complexity also increases. As a compromise, three time-varying amplitudes with one quadratic and two linear functions for the sinusoidal model are used for the proposed speech coding system.

4. Estimation of Dominant Frequencies $\omega_{k,n}$

The estimation of the dominant frequencies for the enhanced sinusoidal model is one of the important issues for the model. The dominant frequencies cannot be simply extracted by picking up the components with relatively larger magnitudes using the DFT, because most of components with higher magnitudes are concentrated on the first and second formats. Using such an approach, it is not possible to reconstruct the components of formats in high frequency bands, because of the lack of dominant components in these bands. Compared with the original spectrum in

Model Parameters	No. of Par.	SNR (dB)	Perceptual Quality
10 sine waves with constant amplitudes	30	7.94	poor
9 sine waves with one linear func.	29	10.17	good
9 sine waves with two linear func.	31	11.40	good
8 sine waves with 3 linear func.	30	11.89	good
7 sine waves with 2 lin. 1 quad.	29	13.20	better
6 sine waves with 1 lin. 2 quad.	28	13.79	better

Table 2 Comparisons Between 6 Time-varying Amplitudes

Fig. 2 , the distortion due to the simple DFT method with high cepstrum distance is apparent, as shown in Fig. 3.

The estimation of the dominant frequency from the pitch and formats, as developed in this paper, is determined by the following steps:

1. Pitch estimation $\omega_{p,n}$;
2. Formants extracted from the LPC spectrum;
3. First, only 1 or 2 peak components in each format are considered to the dominant frequencies $\{ \omega_{k,n} \}$;
4. The remaining dominant frequencies are defined by extracting those larger magnitudes of the spectrum. The total number of dominant frequencies is limited to 20;
5. The intervals of the consecutive dominant frequencies are checked. If $\omega_{k,n} - \omega_{(k-1),n} > 7\omega_{p,n}$, additional dominant frequencies are inserted into the gaps.

The pitch extraction algorithm, referred to as parallel processing in the frequency domain [4], has been used to extract pitch frequency with pitch errors smaller than 0.5%. The errors mainly occur in low pitch and transition segments. The purpose of Steps (3) and (5) is to make the distribution of the dominant components more uniform in the spectrum. This allows for the recovery of the original spectrum within the gaps using the linear prediction of the dominant components and harmonics . The spectrum of the synthesized speech is shown in Fig. 4. The cepstral distance improves by 1.39 dB, compared with the simple DFT method.

5. Parameters Quantization

The total number of parameters to be quantized is equal to 65 for the enhanced sinusoidal model. This count includes 20 sinusoidal components, with amplitudes, dominant frequencies, phases, pitch, and three time-varying amplitudes. Assuming the length of the speech analysis frame is 22.5 ms, (180 samples at a 8 kHz sampling rate), the total number of available bits is equal to 108 bits/frame for a 4.8 kbits/s speech coding system. Since the number of bits is rather low, vector quantization is used for the dominant frequencies, amplitudes and phases. However, the pitch, the locations of the three sinusoids with time-varying amplitudes, and the phase prediction residuals are quantized using scalar quantization. The bit allocations for the speech coding system using the enhanced sinusoidal model at 4.8 kbits/s are listed in Table 3. (Notations: SQ-Scalar Quantization; VQ-Vector Quantization; GSPVQ-Gain-Shape Product VQ)

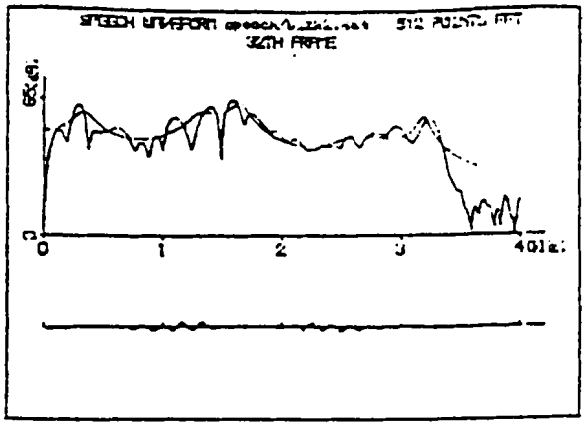

Fig. 2 Original Speech

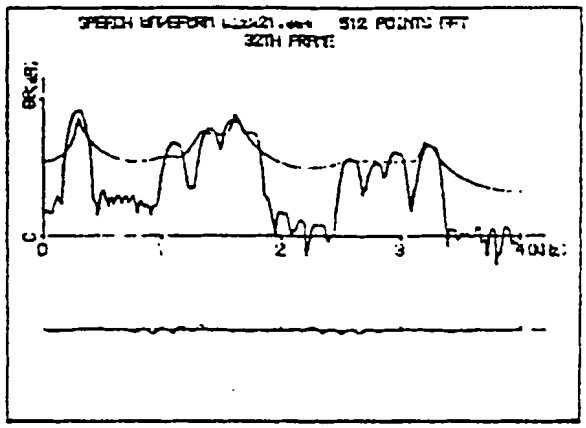

Fig. 3 Synthesis Speech with Estimation of Dominant Frequency by DFT

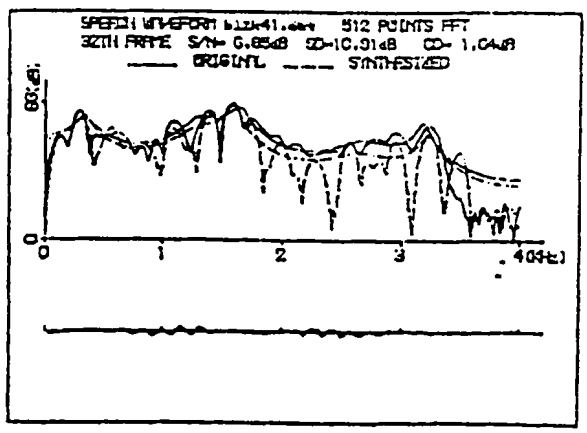

Fig. 4 Synthesis Speech with Estimation of Dominant Frequency by Formats and Pitch

Parameters	bits
Synchronization	2
Pitch	5 SQ
Dominant Freq.	10 VQ
locations for Three TV com.	13 SQ
Amplitudes of 2 linear 1 quad.	13 GSPVQ
Amplitudes of constant ampl.	26 GSPVQ
Phase prediction	8 SQ
3 dominant com.	8 VQ
Phase residuals	22 SQ
total	108

Table 3 Bit Allocations

The reconstructed speech waveforms and spectrums of the proposed speech coding system are shown in Fig. 5 and 6. The solid curves represent the original speech signal and the dashed curves stand for the synthesized signal. The cepstrum distance is 2.62 dB.

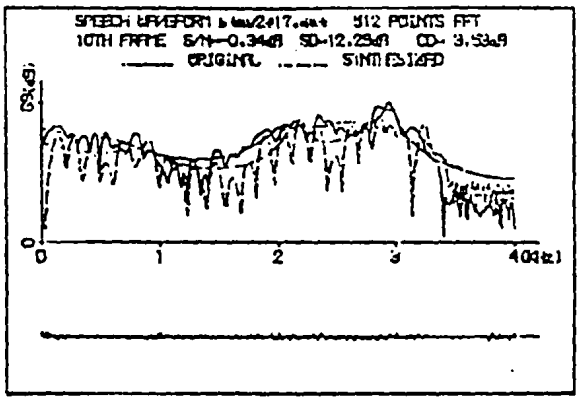

Fig. 5 The Synthesis Waveforms and Spectrum for a Female Speech

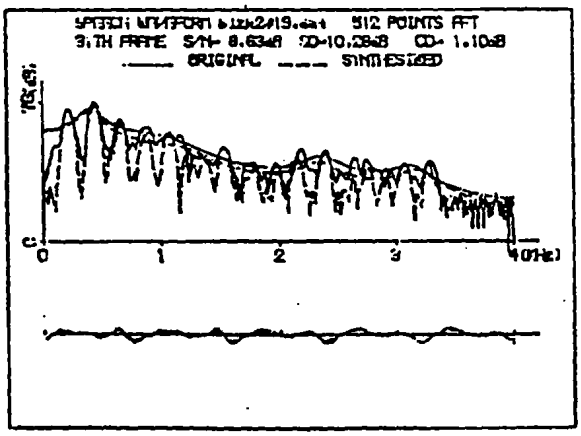

Fig. 6 The Synthesis Waveforms and Spectrum for a Male Speech

6. Conclusions

The speech coding system using an enhanced sinusoidal model at 4.8 kbits/s has been presented. The proposed sinusoidal model describes speech signals by a few dominant frequency sinusoids with three time-varying amplitudes and harmonic components. It has reduced the number of the parameters, especially, in the unvoiced and transition segments by 2/3 while maintaining high quality. The algorithm for extracting the dominant frequencies from the formats and pitch is better than a simple DFT method. The choice of three time-varying amplitudes with two linear and one quadratic functions is a good trade-off between speech quality and complexity. Phase prediction can also be applied to reduce the parameters for transmission. The objective tests with cepstrum distance of 2.62 dB for several phonetically balanced sentences and informal listening comparisons have shown that the speech coding system using an enhanced sinusoidal model can reach to good quality at 4.8 kbits/s.

References

1. R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans., Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug. 1986.
2. R. J. McAulay and T. Champion, "Improved interoperable 2.4 kb/s LPC using sinusoidal transform coder techniques, "*Proc. Int. Conf. Acoust., Speech, Signal Processing* (Albuquerque, NM), April, 1990, pp. 641–643.
3. J. S. Marques, L. B. Almeida and J. M. Tribolet, "Harmonic coding at 4.8 kb/s," *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Albuquerque, NM), April, 1990, pp. 17–20.
4. Liu Jia, Qian Yasheng and Feng Chongxi, "A new hybrid speech model with a time domain perceptual weighting function at 9.6 kbits/s," *Int. Symp. on Comp. Archit. and Digital Signal Processing*, (Hong Kong), Oct. 1989.