ENHANCED HARMONIC CODING OF SPEECH WITH FREQUENCY DOMAIN TRANSITION MODELING

Chunyan Li and Vladimir Cuperman

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106

ABSTRACT

A major source of audible distortion in current low-bit-rate harmonic speech coding algorithms is the ineffective modeling of the transitional speech signals such as onsets, plosives etc.. A new method of modeling transitional speech based on a frequency domain approach is introduced in this paper. The approach uses a modified harmonic model able to produce non-periodic pulse sequences in conjunction with a closed-loop analysis-by-synthesis scheme for parameter estimation and quantization. The structure of a speech coding system based on this model is outlined. The proposed approach is shown to give better performance than transition encoding based on a standard CELP algorithm at rates of 4-8kb/s.

1. INTRODUCTION

In recent years CELP algorithms have been dominant in speech coding at rates above 4kb/s. At lower rates, however, CELP systems suffer from large amounts of quantization noise due to the fact that there are not enough bits to accurately encode the details of the waveform. An alternative approach in low-bit-rate speech coding is frequency domain based harmonic coding (sinusoidal coding) employed in coders like Sinusoidal Transform Coding (STC) [1], Multiband Excitation Coding (MBE) [2], Time Frequency Interpolation (TFI) [3], Spectral Excitation Coding [4], and Hybrid Speech Coding [5].

Harmonic coders are well-suited for the reconstruction of quasi-periodic signals typical of voiced speech. An analysis in [1] using the Karhunen-Loeve expansion for the noise-like signals shows that the harmonic model is also valid for unvoiced speech provided that fundamental frequency used in spectral sampling is less than approximately 100Hz. However, the harmonic models are ineffective for representing speech in transition regions such as voicing onsets, plosives and nonperiodic pulses. Experimental evidence shows that the reconstructed speech has a "buzzy" quality if the standard harmonic models are used for encoding transitional speech.

In this paper we introduce a generalized harmonic model which can reproduce non-periodic sequences of pulses typical of transitional speech. Experimental evidence is presented to show that the new model improves the quality of harmonic coding and results in better performance on transitional speech than CELP type coders.

*This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Advanced Computer Comm., Cisco Systems, DSP Group, DSP Software Eng., Fujitsu, General Electric Company, Hughes Electronics, Intel, Nokia Mobile Phones, Qualcomm, Rockwell International, and Texas Instruments.

2. HARMONIC CODING OF VOICED AND UNVOICED SPEECH

Voiced and unvoiced speech can be synthesized using the harmonic model:

$$\hat{e}(n) = \sum_{k=1}^{L} A_k(n)\cos(\phi_k(n)) \qquad (1)$$

where $\{A_k\}$are samples of the magnitude spectrum at multiples of the fundamental frequency ω_0, and $\{\phi_k\}$the corresponding phase. This harmonic model has been applied to the speech signal in STC and MBE [1, 2] and to the LP residual in TFI, SEC, and Hybrid Coding [3-5]. In this paper, the model will be applied to the speech LP residual.

For voiced speech, the model is based on the assumption that the perceptually important information resides mainly in the harmonic samples of the pitch frequency. At low rates, the phase is reconstructed from the transmitted pitch values using a quadratic model which assumes linear pitch variation:

$$\phi_k^i(n) = k\omega_0^{i-1}n + \frac{k(\omega_0^i - \omega_0^{i-1})}{2N}n^2 + \varphi_k \qquad (2)$$

where:

- ω_0^{i-1}, ω_0^i are the pitch frequency values for the *i-1th* and the *ith* frame respectively
- N is the frame size in samples
- φ_k is zero for harmonics below a threshold frequency called "voicing" and a random variable uniformly distributed in $[-\pi, \pi]$ for harmonics above the voicing

For unvoiced speech, the magnitude spectrum is sampled at 100 Hz and a uniformly distributed random phase is applied to each frequency component.

3. A FREQUENCY DOMAIN MODEL FOR TRANSITIONAL SPEECH

In order to derive a frequency domain model for transitional speech, we start from the observation that for $\omega_0^i = \omega_0^{i-1} = 2\pi / N$, the signal given by (1, 2) will generate exactly one pulse in a frame. This pulse will be positioned at $n=0$ and for $\varphi_k = 0$ will have even symmetry (note that A_k are all positive). A different shape of the pulse can be obtained by using in (2) a constant phase $\varphi_k = \varphi$ (odd symmetry will result for $\varphi = -\pi/2$). This is illustrated in Fig. 1 (a) for $\varphi = \pi/3$. Next, the pulse can be moved to any position n_0 in the frame by replacing n by $n-n_0$ as shown in Fig 1 (b). Finally, a number of non-periodic pulses of different amplitudes can be obtained by

0-7803-4428-6/98 $10.00 © 1998 IEEE

summing scaled versions of the Fig 1 (b) signal as shown in Fig. 1 (c).

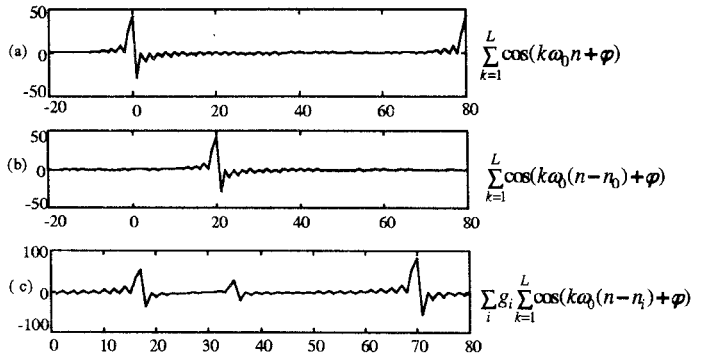

Fig. 1. Sequence of non-periodical pulses generated by the model

Based on the above considerations, we propose to model the transitional residual signal by a generalized sinusoidal model, in which excitation of the *ith* frame or subframe can be synthesized by

$$\hat{e}(n) = \sum_{j=0}^{M-1} g_j \sum_{k=1}^{L} A_k(n)\cos\theta_k(n,n_j) \qquad (3)$$

with the phase term θ_k given by

$$\theta_k(n,n_j) = 2\pi k(n-n_j)/N + \psi_k \qquad (4)$$

where $\{n_j\}$ are the shift parameters representing pulse occurrence times, and ψ_k is a phase vector which affects the pulse shapes.

We assume that the spectral magnitude changes slowly during a frame (10ms in our simulation) so that it is reasonable for all the pulses to use the same spectral envelope parameters $\{A_k\}$ but with different gains $\{g_j\}$. Note that ψ_k in (3) has completely different significance of φ_k in (2): the former is a phase vector which will be chosen to match the model generated pulse shape to the original speech, while the latter is a random phase component added in the model of the "unvoiced" spectrum regions. The change of notation was introduced to emphasize this difference. In examples of Fig. 1, the vector ψ_k was replaced by a constant $\varphi = \pi/3$.

The proposed model preserves time domain information which is important in the perception of the transitional speech by using as parameter pulse occurrence times $\{n_j\}$, while the pulse shape is represented by $\{A_k\}$ and $\{\psi_k\}$. Transition modeling is done in frequency domain, avoiding the use of two completely different coding strategies such as harmonic coding for voiced sounds and a CELP type scheme for transitions [5]. Finally, as shown in the next section, the new model is amenable to a closed-loop analysis-by-synthesis procedure for parameter estimation, which can further improve the robustness.

4. SYSTEM OVERVIEW

Figure 2 shows a block diagram of a harmonic speech coding system based on the model we presented in Section 3. The system is similar to that of [5] except for the transition coding. Once each 10ms frame, the speech spectral envelope is

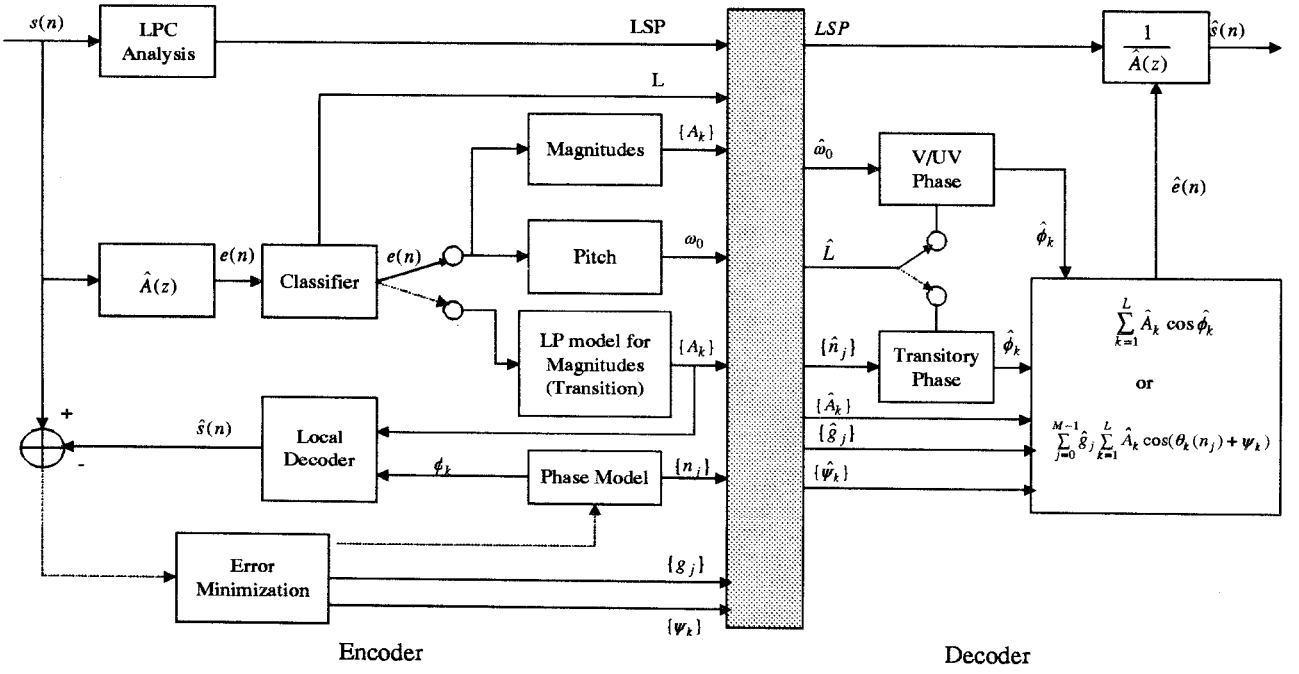

Fig. 2. System block diagram

estimated using a 10*th* order LPC analysis. The coefficients are converted to Line Spectral Pairs (LSP) and quantized once per frame using predictive multistage VQ. The quantized coefficients are then transformed back into LPCs and used in the short-term filter which computes the residual signal. The filter coefficients are updated using LSP interpolation every 5ms. Classification, pitch and voicing are updated for each 10ms frame.

The residual signal is analyzed over 10ms frame in order to estimate the parameters used in the synthesis model. For the voiced speech signal, parameters $\{A_k\}$ are acquired by sampling the magnitude spectrum which is obtained by a windowed DFT. For the unvoiced speech, a dense uniform-interval frequency magnitude sampling is used to obtain the parameters $\{A_k\}$. Voiced and unvoiced excitation signals are both synthesized using (1) with the level of phase randomness controlled by the voicing information.

For the transitional speech, an analysis-by-synthesis scheme is adopted to estimate magnitudes$\{A_k\}$, shifting parameters $\{n_j\}$, the phase vector $\{\psi_k\}$, and gains $\{g_j\}$. Parameter estimation and quantization will be described in detail in the next section. A local synthesizer based on (3) is included in the encoder to perform closed-loop parameter estimation and quantization.

The excitation signal synthesis uses the sinusoidal model given by either (1) or (3) depending on the classification. The reconstructed excitation signal is passed through the inverse short-term filter to obtain the reconstructed speech $\hat{s}(n)$.

5. PARAMETER ESTIMATION AND QUANTIZATION

Experimental evidence shows that a coarse open-loop quantization of the harmonic magnitudes $\{A_k\}$ is acceptable for transition coding when using the model of (3). For each transition speech frame, a *10th* order all-pole model is derived from a windowed DFT of the excitation signal. The magnitudes $\{A_k\}$ are obtained by sampling this spectral envelope at equal intervals. The all-pole model coefficients are converted to LSP domain and vector quantized using a small number of bits.

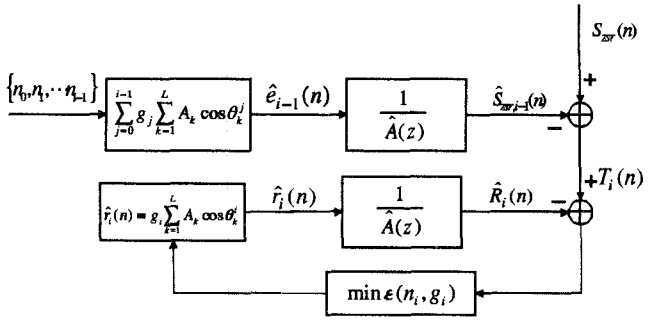

Fig. 3. Combined closed-loop search procedure

Figure 3 shows a block diagram of the closed-loop analysis-by-synthesis procedure for determining shifts $\{n_j\}$ and gains $\{g_j\}$.

The objective of the estimation procedure is to match the original speech signal by minimizing the mean squared error :

$$E = \sum_{n=0}^{N-1} (S_{zsr}(n) - h(n) * \sum_{j=0}^{M-1} g_j \sum_{k=1}^{L} A_k \cos\theta_k(n, n_j))^2 \qquad (5)$$

where the target vector $S_{zsr}(n)$ is obtained by subtracting the zero input response of the short-term synthesis filter from the original speech signal, $h(n)$ is the impulse response of the short-term synthesis filter, and * stands for convolution. The minimization is done using a sub-optimal sequential search. The contribution of the pulse i, $\hat{R}_i(n)$, is given by

$$\hat{R}_i(n) = g_i \sum_{k=1}^{L} A_k \cos\theta_k(n, n_i) * h(n) \qquad (6)$$

Suppose now that the parameters of the first *i-1* pulses have been determined previously, resulting in a sequential approximation to $S_{zsr}(n)$ which can be expressed as

$$\hat{S}_{zsr,i-1}(n) = \sum_{k=1}^{i-1} \hat{R}_k(n) \qquad (7)$$

The target vector for the estimation of the *ith* pulse parameters is then given by

$$T_i(n) = S_{zsr}(n) - \hat{S}_{zsrr,i-1}(n) \qquad (8)$$

The parameter n_i is found by an exhaustive search whereby for each value of n_i the optimal value of the corresponding gain, g_i, is computed to minimize,

$$E_i = \sum_{n=0}^{N-1} (T_i(n) - g_i h(n) * \sum_{k=1}^{L} A_k \cos\theta_k(n, n_i))^2 \qquad (9)$$

After the last stage, the gains $\{g_j\}$ are re-optimized and vector quantized in log-energy domain by mean removed MSVQ. The gain re-optimization is done by solving a system of linear equations which is obtained by minimizing the error between $S_{zsr}(n)$ and the reconstructed signal :

$$E(g_0, g_1, \ldots, g_{M-1}) = \sum_{n=0}^{N-1} [S_{zsr}(n) - \sum_{j=0}^{M-1} \hat{R}_j(n)]^2 \qquad (10)$$

System performance could be further improved at the expense of increased complexity by using closed-loop gain vector quantization.

Based on the approach described above, we implemented transition encoding at rates of 7.4kb/s and 4.7kb/s. At the lower rate, the phase ψ_k in (3) is discarded (set to zero). At the higher rate, the phase vector $\{\psi_k\}$ is quantized using a mean-shape decomposition: $\psi_k = \psi'_k + \psi$ where ψ is a constant phase in the range -0.5π to 0.5π which is scalar quantized using 9 bits. The shape phase vector $\{\psi'_k\}$ is selected from a codebook composed of 1024 phase vectors. These vectors are generated as uniformly distributed random variables in a narrow dynamic range. In our tests we found that the range -0.1π to 0.1π gave good performance. ψ and $\{\psi'_k\}$ for each frame are selected sequentially by analysis-by-synthesis minimizing the error criterion (5).

6. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed model, the system presented in Section 4 was implemented at rates of 4.7kb/s and

7.4kb/s. The parameter quantization is done as described in Section 5. The bit allocations for each rate are given in Table 1 (combinatorial coding was assumed for shifts).

Table 1. Bit Allocation for the transitional speech frame

LPC	10	16
Magnitudes	4	4
Shifts	19	19
Gains	10	12
Signs	3	3
Dispersion Phase	0	19
Classifier	1	1
Total	47	74
Bitrate	4.7kb/s	7.4kb/s

We compared the proposed coders at both rates with the G.729 standard on a data-base of 2,443 transition frames selected by classification from 22,000 frames of speech. The G.729 coder was run on the entire speech data base to obtain for each transition frame same initial conditions that would result if all the speech material would have been encoded by G.729. For the proposed coders, the initial conditions are irrelevant and only transition frames need to be encoded. The average segmental SNR for these three coders are shown in table 2. The comparison results for two particular sentences, a female and a male, are shown in Fig. 4 and Fig. 5 respectively. Results in Table 2 show that the proposed coder achieves better objective performance than G.729 even at the lower rate. This result may seem surprising, however we have to consider the fact that G.729 was not designed for transition coding and particularly the adaptive codebook may be inefficient on transitional speech.

Table 2. Average segmental SNR of the transitional speech

Coders	Average seg. SNR (dB)
Proposed 7.4kb/s	7.102 dB
Proposed 4.7kb/s	5.332 dB
G.729	4.011 dB

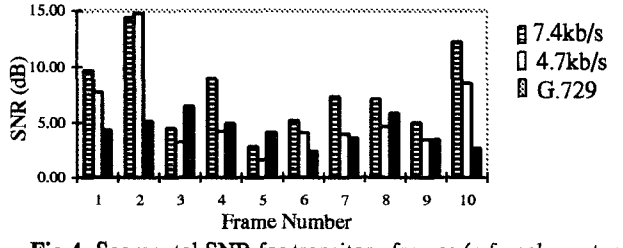

Fig.4. Segmental SNR for transitory frames (a female sentence)

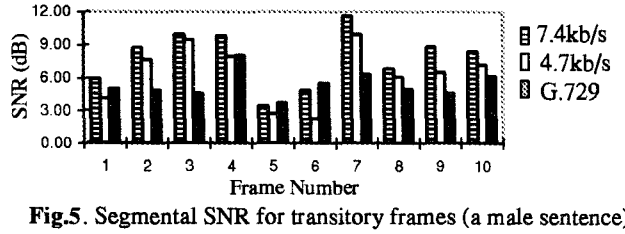

Fig.5. Segmental SNR for transitory frames (a male sentence)

To evaluate the subjective performance, a preference listening test was performed. In the test speech files, voiced/unvoiced frames were coded by the harmonic coding method described in Section 3 (see also [5]). The transition frames were coded by the proposed coder (at 4.7 and 7.4 kb/s), the G.729 coder, or the standard harmonic coder for comparison. The results shown in Table 3 indicate that the proposed approach outperforms the standard G.729 codec (for transition coding) at the same rate (or allows a significant rate reduction for similar quality) and improves the performance of the harmonic codec.

Table 3. Preference Test Results

System 1 : 2	Pref. 1	Pref. 2	No pref.
4.7k : G.729	45%	48.75%	6.25%
7.4k : G.729	66.25%	20%	13.75%
4.7k : HC*	55%	35%	10%

*HC : Harmonic Coder

REFERENCE

[1] R.McAulay and T. Quatieri, "Sinusoidal Coding" in *Speech Coding and Synthesis*, W.B.Kleijn and K.K.Paliwal, Eds., Chapter 4, Elsevier, 1995

[2] Daniel W. Griffin and Jae S. Lim, "Multiband Excitation Vocoder" *IEEE Trans. On ASSP*, Vol. 36, No. 8, August 1988

[3] Yair Shoham, "High-Quality Speech Coding at 2.4 to 4.0 KBPS Based on Time-Frequency Interpolation", *ICASSP'93*, pp. II-167

[4] P. Lupini and V. Cuperman, "Spectral Excitation Coding of Speech at 2.4 kb/s", in *Proc. ICASSP*, 1995.

[5] E. Shlomot, V. Cuperman and A. Gersho, "Hybrid Coding of Speech at 4kbps", in *Proc. of IEEE Workshop on Speech Coding*, pp.37-38, 1997.