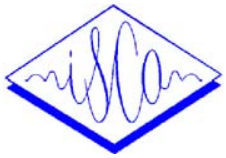

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ANALYSIS-BY-SYNTHESIS LOW-RATE MULTIMODE HARMONIC SPEECH CODING

Chunyan Li, Allen Gersho and Vladimir Cuperman

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

10.21437/Eurospeech.1999-329

ABSTRACT

This paper presents an analysis-by-synthesis multimode harmonic coder (AbS-MHC) that employs new techniques to improve both the speech model accuracy and the parameter estimation robustness in the low rate harmonic coding framework. To improve the speech model accuracy, an enhanced frequency domain transition model is used in conjunction with the sinusoidal model based harmonic coding of voiced/unvoiced speech signals. To achieve robust parameter estimation, a generalized analysis-by-synthesis parameter estimation scheme in the harmonic coding framework is proposed. This scheme uses a time scale signal modification technique to allow for waveform matching in harmonic coding. This concept is demonstrated in our AbS-MHC coder with a specific method for efficient closed-loop pitch estimation and speech classification. The speech quality of the unquantized AbS-MHC coder is better than the 6.3 kbps G.723 quality.

1. INTRODUCTION

High quality speech coding at 4 kbps and below is of major interest in speech coding research. Waveform coders such as CELP [1] are able to produce high quality speech at bit rates as low as 6.3 kbps. As the bit rate is reduced to 4 kbps and below, waveform coding systems suffer from large amounts of quantization noise because there are not enough bits to accurately encode the details of the waveform. On the other hand, parametric coders (also called vocoders) [2], [3], [4], [5] do not attempt to reproduce a waveform similar to the original. Instead, those coders try to find a parametric representation of the speech signal that captures its perceptually essential characteristics. Particularly, some vocoders exploit the perceptually important information that can be usually represented as the harmonically related line structure of the speech spectrum. Those vocoders that make use of the harmonic structure of speech spectrum are usually referred to as "harmonic coders". It was demonstrated that retaining only the spectral harmonic magnitudes and using a synthetic harmonic phase is sufficient for high quality reproduction of voiced speech [6]. Therefore, harmonic coders are attractive methods to obtain high quality reconstructed speech at low bit rates.

Though harmonic coders currently have been widely used for low-bit rate speech coding, the reproduced speech quality of harmonic coders is limited by the accuracy of the harmonic model. What is more important, since harmonic coders heavily rely on correct parameter estimates, the model distortion cannot be easily controlled due to the open-loop parameter estimation typical of harmonic coders. Therefore, in designing the proposed analysis-by-synthesis multimode harmonic coder (AbS-MHC), we try to overcome the above limitations of harmonic coders by employing new techniques to improve both the speech model accuracy and the parameter estimation robustness. To improve the speech model accuracy, an enhanced frequency domain transition model is used in conjunction with the sinusoidal model based harmonic coding of voiced/unvoiced speech signals. To achieve robust parameter estimation, a generalized analysis-by-synthesis parameter estimation scheme in the harmonic coding framework is proposed. This scheme employs a time scale signal modification technique to allow for waveform matching in harmonic coding. This concept is demonstrated in our AbS-MHC coder with a specific method for efficient closed-loop pitch estimation and speech classification.

Subjective test results show that the speech quality of the unquantized AbS-MHC coder exceeds that of G.723 coder at 6.3 kbps. Initial efforts towards a fully quantized 4 kbps coder have produced the speech quality which is comparable to G.723 coder operating at 5.3 kbps. Particularly, for the modified IRS filtered speech, the speech quality of the 4 kbps AbS-MHC coder is better than that of G.723 at 5.3 kbps.

2. NEW TECHNIQUES IN THE PROPOSED CODER

2.1. Improved Speech Model for Transitions

In harmonic coding, voiced and unvoiced speech can be synthesized using the harmonic model:

$$\hat{e}(n) = \sum_k A_k(n)\cos\theta_k(n), \tag{1}$$

where A_k are samples of the magnitude spectrum at multiples of the fundamental frequency, and θ_k the corresponding phase. 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 value using a quadratic model which assumes linear pitch variation:

$$\theta_k(n) = k[\frac{2\pi}{F_s}({f_0}^{(i-1)}n + \frac{{f_0}^{i} - {f_0}^{(i-1)}}{2N}n^2)] + \varphi_k, \tag{2}$$

where $f^{(i-1)}$, f^i are pitch frequency values for the $i-1$ th and the i th frame respectively, F_s is the sampling rate, 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 frequency. For unvoiced speech, the magnitude spectrum is sampled at 100 Hz and a uniformly distributed random phase is applied to each frequency component.

Though the harmonic model is well-suited for the reconstruction of voiced and unvoiced signals, it is ineffective for representing the transition speech signals such as voicing onsets, plosives and nonperiodic pulses. Therefore, in the AbS-MHC coder, an enhanced frequency domain transition model is used in conjunction with the conventional harmonic model based harmonic coding of voiced/unvoiced speech signals. In this new transition model, the transition LP residual signal is modeled by a generalized sinusoidal model, in which the excitation of the frame or subframe can be synthesized by

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

with the phase θ_k given by

$$\theta_k(n, n_j) = 2k\pi(n - n_j)/N + \varphi_k, \tag{4}$$

where $\{n_j\}$ are the shift parameters representing pulse occurrence times, and $\{\varphi_k\}$ is a phase vector which affects the pulse shapes. We assume that the spectral magnitude vector $\{A_k\}$ changes slowly during a frame (10ms in our coder) so that it is reasonable for all the pulses to use the same spectral envelope parameters but with different gains $\{g_i\}$.

This new transition model is able to produce non-periodic sequences of pulses typical of transition speech signal. In this model, pulse occurrence information, which is the most perceptually important information in the transition frame, is represented by parameters $\{n_j\}$. Most of the available bits are used to accurately quantize these parameters. The dispersion phase $\{\varphi_k\}$ and spectral magnitude $\{A_k\}$ together represent the pulse shapes. Experimental evidence shows that coarse quantizations of A_k and φ_k are perceptually acceptable. In our experiment, the dispersion phase vector is replaced by a scalar variable and is quantized by a simple uniform scalar quantizer. The spectral magnitude vector is approximated by the spectral envelope which can be derived from a 10-th order all-pole model. The all-pole model parameters are vector quantized in the LSF domain. This transition model is amenable to a closed-loop analysis-by-synthesis procedure for parameter estimation and quantization.

2.2. Analysis-by-Synthesis Parameter Estimation in the Harmonic Coding Framework

Since harmonic coders belong to the parametric coding category, they can be successful only if the model parameters are estimated accurately. In other words, the estimation errors for the model parameters would result in significant degradation of the speech quality. One solution for improving the accuracy and robustness of parameter estimations is to use time domain closed-loop analysis-by-synthesis techniques. However, when low-rate harmonic coders

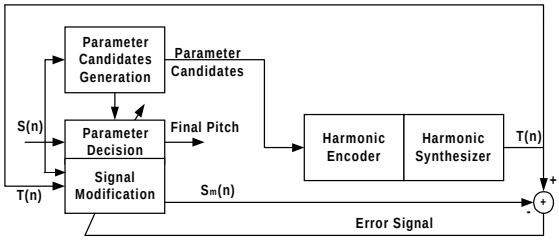

Figure 1: Analysis-by-synthesis parameter estimation in harmonic coding

are used to synthesize speech, no phase information is transmitted, which results in a loss of time alignment between the original speech and the synthesized speech. This loss of time alignment makes it difficult for the harmonic coder to perform waveform matching, and interferes with time domain closed-loop parameter estimation. We have found, however, that when a suitable time-scale modification is applied to the original speech signal, the harmonic coders can benefit from waveform matching. The concept of performing analysis-by-synthesis parameter estimation in the harmonic coding framework is illustrated in Figure 1. Several candidates of parameter estimates are first open-loop generated. For each candidate, the signal modification module performs time scale signal modification on the original speech signal or the original LP residual signal $S(n)$ under the constraint that the modified signal $S_m(n)$ will give the identical perceptual quality as the original signal. The reference signal or the target signal for the signal modification module, $T(n)$, is generated by the harmonic synthesizer based on the current parameter candidate. If the current parameter candidate is a good estimate, it will be easy for the signal modifier to align the original signal to the target signal while the perceptual quality is still preserved. Therefore, the error signal between the target signal and the modified signal will be small. On the other hand, an incorrect parameter estimate will make the signal modification difficult under the constraint that the perceptual quality should be identical after the signal modification. Therefore, the modified signal will be no longer aligned with the target signal. That will lead to a large error signal between the modified signal and the target signal. This error signal is fed back to both the signal modification procedure and the parameter estimation module; it is then used to adaptively control the final parameter decision.

We applied the above generalized concept of time domain AbS parameter estimation to the proposed AbS-MHC coder with a specific algorithm for the time domain closed-loop pitch estimation and classification. The algorithm has three stages. The first stage pre-classifies the input speech into one of two categories: the first category includes unvoiced speech and silence; while the second includes voiced speech and transition speech. This stage also generates 5 pitch candidates corresponding to the local maxima of the autocorrelation function of the low-pass filtered input LP residual signal. The second stage is performed only on the voiced speech and the transition speech to perform the voiced/transition speech classification and determine the pitch. At the last stage, a pitch refinment and harmonic bandwidth estimation procedure is performed on subframes which are declared as voiced. This procedure is similar to that described in [7]. At the second stage, the time-

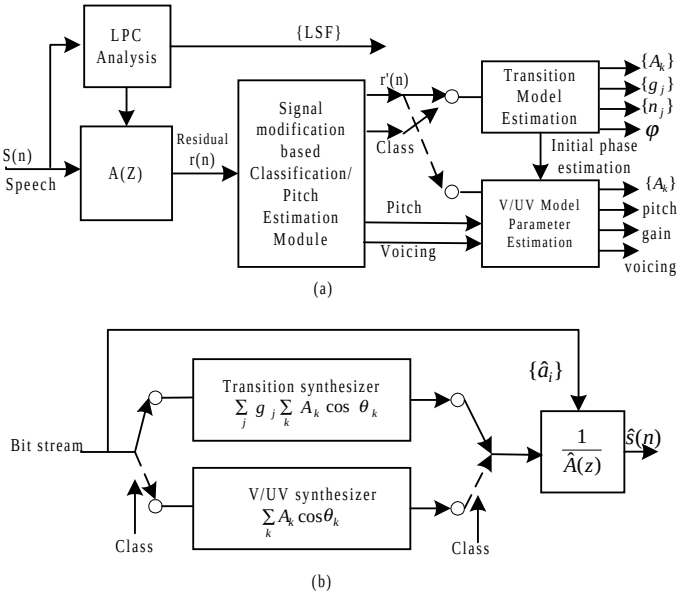

Figure 2: Simplified block diagram of AbS-MHC speech codec (a) encoder (b) decoder.

domain closed-loop pitch estimation based on Figure 1 is used for pitch estimation. The signal modification procedure we used is similar to that used in the EVRC coder [8]. The closed-loop information from the time scale signal modification is also used to help making the voiced/transition decision. This is based on the assumption that if a voiced subframe is claimed, we should be able to find a reasonable pitch value for that subframe, which will result in a good alignment between the modified signal and the synthetic signal. The classification decision between the transition speech and the voiced speech is then made according to the following parameters: normalized signal energy, pitch variation across the subframe, energy variation across the subframe, the time domain autocorrelation of the pitch lag, and the normalized correlation between the modified residual signal and the synthetic residual signal.

Experimental results show that the new algorithm for time domain closed-loop pitch estimation based on time scale signal modification significantly reduces gross pitch errors compared to a conventional time domain pitch estimator based on the autocorrelation function. For example, we calculated the percentage of the pitch outliers which have normalized pitch error larger than 10% for both clean and noisy speech sentences. Compared to the conventional time domain pitch estimator, the closed-loop algorithm reduces the outliers from 7.3% to 2.9% for clean speech, from 7.5% to 3.6% under office noise condition, from 8.6% to 3.8% under harmonic noise condition, and from 8.4% to 3.8% under babble noise conditions.

3. ABS-MHC SPEECH CODER

New techniques described in the above section are applied to develop the analysis-by-synthesis multimode harmonic speech coder. The simplified block diagram of the AbS-MHC codec is shown in Figure 2. In the encoder, a Linear Prediction Coding (LPC) module is used to obtain the LP residual signal, which is the target signal for the AbS-MHC coder. A signal modification based classification/pitch estimation/voicing detection module described in Section 2.2 operates on the LP residual signal and classifies the input frame into one of three classes – voiced speech, unvoiced speech, and transition speech. The encoder is set to a particular encoding mode according to the classification. For voiced/unvoiced frames, the speech model (1) is used and model parameters including harmonic magnitude $\{A_k\}$, pitch, voicing frequency and gain are estimated and transmitted to the decoder. For transition frames, the transition model (3) described in Section 2.1 is used and the corresponding parameters are extracted and transmitted to the decoder. According to the received class information, the decoder is set to a particular decoding mode. For each mode, an appropriate excitation synthesizer is used to synthesize the LP excitation signal based on the decoded parameters. The reconstructed excitation signal is then passed through an LP synthesis filter to generate the reconstructed speech signal.

The use of different coding techniques for the voiced mode and the transition mode requires waveform synchronization at the boundary between these modes. When going from a voiced frame to a transition frame, the drift between the original signal and the synthetic one is measured by the time scale signal modification procedure which is performed in the classification/pitch estimation module. This drift is kept as the accumulated shift parameter τ_{acc}. The synchronization can be done by simply shifting the following original transition frames by τ_{acc} and encoding the shifted residual signal. τ_{acc} will be carried over all the following consecutive transition frames, and can be reset when a unvoiced frame is encountered. When a transition frame is followed by a voiced frame, the initial linear phase is estimated by maximizing the correlation of the shifted voiced frame with the transition frame [7].

The AbS-MHC coder operates on speech sampled at the rate of 8 kHz. The frame size is 20 ms and the subframe size is 10 ms. The look-ahead for the harmonic analysis is 20 ms and the look-ahead for the LP analysis is 10 ms. A 10-th order LPC analysis is performed once per frame, but all other processing is performed on each subframe.

3.1. Initial 4 kbps quantization for Voiced/Unvoiced Modes

The voiced/unvoiced LP residual signal can be represented by the unified harmonic model (1). For the unvoiced speech, the pitch frequency is set to 100 Hz and the voicing frequency is set to 0. The bit allocation for 4 kbps coding of voiced/unvoiced modes is given in Table 1. The 10 LSF coefficients are vector quantized

Parameters	1st subframe	2nd subframe	frame
LSFs			24
class			1
pitch	0	7	7
harmonic magnitudes	14	14	28
voicing	0	6	6
gain	7	7	14
total			80

Table 1: Bit allocation for voiced/unvoiced modes at 4 kbps

by a 3-stage MSVQ with 8 bits per stage. The pitch for the 2nd subframe is searched from 20 to 144 samples and quantized using 7 bits. The pitch of the 1st subframe is obtained by linear interpolation. The variable-dimension harmonic magnitude vector is quantized by DCT-II transform based weighted non-squared transform VQ (WNSTVQ) scheme [9] using a 14-bit MSVQ quantizer. The voicing frequency is quantized using 6 bits for the 2nd subframe and is not transmitted for the 1st subframe; the latter value is obtained by linearly interpolating the adjacent voicing cut-off frequencies. The gain is scalar quantized in the logarithm energy domain using 7 bits for each subframe.

3.2. Initial 4 kbps quantization for Transition Mode

The transition frame coding is based on the proposed transition model (3). The bit allocation for the 4 kbps coding of transition frame is given by Table 2. The 10 LSF coefficients are vector

Parameters	subframe(10 ms)	frame (20 ms)
LSFs		18
class		1
$\{A_k\}$	5	10
dispersion phase φ	2	4
shifts $\{n_j\}$	$5+4+5$	28
signs	3	6
gains $\{g_j\}$	6	12
total		79

Table 2: Bit allocation for the transition mode at 4 kbps

quantized by a 3-stage MSVQ with 6 bits each stage. Spectral envelope $\{A_k\}$ is modeled by a 12-th order all-pole model and the model parameters are quantized by a 5-bit vector quantizer. The dispersion phase vector is replaced by a scalar which is quantized by a 2-bit scalar quantizer. In each subframe (10ms), 80 positions are divided into 3 grid tracks and 1 pulse will be found in each track. Pulse positions are encoded on a grid using 14 bits. Each pulse is assigned a different sign indicated by 1 bit per pulse. The 3 dimension pulse gain vector is vector quantized by a mean removed VQ using 6 bits per subframe.

4. SUBJECTIVE TEST RESULTS

To evaluate the performance of the proposed AbS-MHC coding algorithm, we ran informal A/B (pairwise) listening tests. In the first test, 11 listeners compared the unquantized AbS-MHC coder (AbS-MHC model) with the G.723 coder operating at 6.3 kbps. Sixteen sentences spoken by 8 male and 8 female speakers were used. Test results given in Table 3 show that the speech quality of the unquantized AbS-MHC coder exceeds that of 6.3 kbps G.723 coder.

In another A/B test, listeners compared the fully quantized 4 kbps AbS-MHC coder with the 5.3 kbps G.723 coder. Thirty two test speech sentences were used including 20 flat speech sentences and 12 modified IRS filtered speech sentences. Test results in Table 4 show that the 4 kbps AbS-MHC coder can achieve speech quality comparable to the 5.3 kbps G.723 coder. Particularly, for the modified IRS filtered speech sentences, the test indicates a

	Pref. G.723 (6.3 kbps)	Pref. AbS-MHC (model)	No pref.
Overall	39.8%	52.2%	8.0%
Female	44.3%	46.6%	9.1%
Male	35.2%	58.0%	6.8%

Table 3: A/B test results between G.723 at 6.3 kbps and AbS-MHC model

preference of 50% for the AbS-MHC coder at 4.0 kbps versus 39% for G.723 coder at 5.3 kbps and 11% for no preference.

	Pref. G.723 (5.3 kbps)	Pref. AbS-MHC (4.0 kbps)	No pref.
Overall	40.2%	45.7%	14.1%
Female	39.8%	49.2%	11.0%
Male	40.6%	42.2%	17.2%

Table 4: A/B test results between 5.3 kbps G.723 coder and 4.0 kbps AbS-MHC coder

5. REFERENCES

[1] B. Atal and J. Schroeder, "Stochastic coding of speech signals at very low bit rates," in *Proceedings of the International Conference on Communications*, pp. 1610–1613, 1984.

[2] D. Griffin and J. Lim, "Multi-band excitation vocoder," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1223–1235, Aug. 1988.

[3] R. McAulay and T. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), Amsterdam: Elsevier Science Publishers, 1995.

[4] W. Kleijn and J. Haagen, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), Amsterdam: Elsevier Science Publishers, 1995.

[5] A. McCree and T. B. III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transaction on Speech and Audio Processing*, vol. 3, pp. 242–250, July 1995.

[6] L. Ameida and J. Tribolet, "Non-stationary spectral modeling of voiced speech," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 31, pp. 664–678, June 1983.

[7] E. Shlomot, V. Cuperman, and A. Gersho, "Combined harmonic and waveform coding of speech at low bit rates," in *Proc. IEEE Intr. Conf. Acoust., Speech, Sig. Process.*, pp. 585–588, 1998.

[8] "TIA/EIA/IS-127, enhanced variable rate codec (EVRC)," in *TIA Draft Standard*, 1996.

[9] C. Li, E. Shlomot, and V. Cuperman, "Quantization of variable dimension spectral vectors," in *Proceedings of the 32nd Asilomar Conference on Signals, Systems & Computers*, pp. 352–356, 1998.