ANALYSIS-BY-SYNTHESIS MULTIMODE HARMONIC SPEECH CODING AT 4 KB/S

Chunyan Li and Vladimir Cuperman

Department of Electrical and Computer Engineering
University of California Santa Barbara, CA 93106
Email: [chunyan, vladimir]@laurel.ece.ucsb.edu

ABSTRACT

This paper presents a 4 kb/s Analysis-by-Synthesis Multimode Harmonic Coder (AbS-MHC). Novel features of this coder include a signal modification technique that allows time-domain analysis-by-synthesis parameter estimation in sinusoidal coding framework, and a frequency-domain transition speech model with improved parameter estimation and quantization schemes. An efficient quantization scheme for harmonic magnitudes based on Weighted Non-Square Transform Vector Quantization (WNSTVQ) is also used. Subjective quality tests indicate that the 4 kb/s AbS-MHC coder outperforms the 5.3 kb/s G.723.1 standard CELP coder and produces speech quality very similar to the 6.3 kb/s G.723.1 coder.

1. INTRODUCTION

Waveform coders such as CELP [1] are able to produce high quality speech at bit rates as low as 6.3 kb/s. At bit rates of 4 kb/s and below, the speech quality in waveform coders degrades 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] employ a parametric representation of the speech signal that captures its perceptually essential characteristics and do not attempt to reproduce a waveform similar to the original.

A particular class of parametric coders exploits the perceptually important information that can be usually represented as the harmonically related line structure of the speech spectrum. These coders represent the signal as a sum of harmonically related sinewaves and are referred to as sinusoidal or harmonic coders.

The sinusoidal model can be applied directly to the speech signal [2], [3] or to the LP residual [6]. In this paper, the sinusoidal model will be applied to the LP residual:

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

where A_k are samples of the magnitude spectrum at multiples of the fundamental frequency, and θ_k the corresponding phases. For voiced speech, the phase is reconstructed from the transmitted pitch values using a quadratic model which assumes linear pitch variation:

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

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 and N is the frame size in samples. The phase term φ_k is set to zero for harmonics below a threshold frequency called "voicing" and to 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. It has been shown that retaining only the spectral harmonic magnitudes and using a synthetic harmonic phase is sufficient for high quality reproduction of voiced speech [7]. Therefore, harmonic coders are attractive methods to obtain high quality reconstructed speech at low bit rates.

Though harmonic coders have been widely used to produce good quality speech at low bit rates, they do not achieve toll quality. This is mainly due to the large modeling distortion for the non-stationary speech segments, and to robustness problems in parameter estimation due to the open-loop estimation typical of harmonic coders.

When low-rate harmonic coders are used to synthesize speech, the absolute phase information is usually not transmitted, which results in a loss of time alignment between the original signal and the synthesized signal. This loss of time alignment makes it difficult for the harmonic coder to perform waveform matching and 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. This concept is employed in our AbS-MHC coder to improve the robustness and accuracy of pitch estimation and classification.

In order to improve the speech model accuracy in non-stationary speech segments, the AbS-MHC coder uses a novel frequency-domain speech model for the transition coding. This model represents time-domain significant events (pulses) by using a generalized sinusoidal model (see Section 3).

Efficient quantization of the variable-dimension harmonic magnitude vector is crucial for achieving high quality reproduced speech in harmonic coding. We developed a quantization scheme for harmonic magnitudes using the DCT-II transform based Weighted

This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, Cisco Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Intel Corp., Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., Rockwell International Corp., Panasonic Technologies, Inc., and Texas Instruments, Inc.

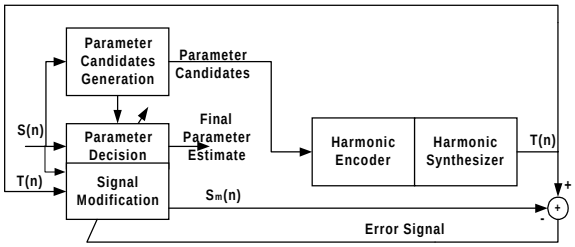

Figure 1: Analysis-by-synthesis parameter estimation in harmonic coding

Non-Square Transform Vector Quantization (WNSTVQ) in conjunction with intra/inter-vector interpolations.

In this paper, we present a 4 kb/s AbS-MHC coder that employs the above new techniques. Subjective quality tests indicate that the 4 kb/s AbS-MHC coder outperforms the 4 kb/s MPEG-4 standard and the 5.3 kb/s G.723.1 CELP coder, and it produces speech quality very similar to the 6.3 kb/s G.723.1 coder.

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 order to improve the robustness of parameter estimation in sinusoidal coders, we introduce a time-scale modification technique that allows closed-loop time-domain parameter estimation. The concept of performing analysis-by-synthesis parameter estimation in the harmonic coding framework is illustrated in Figure 1. Several candidates of the parameter estimate are first generated using an open-loop configuration. 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 perceptual quality identical to 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 perceptual quality preservation constraint. 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 decision module; it is then used to adaptively control the final parameter decision.

We applied the above concept of time domain AbS parameter estimation to the 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; the second includes voiced speech and transition speech. This stage also generates several pitch candidates based on both frequency-domain pitch estimation method [8] and time-domain pitch estimation method. The time-domain pitch candidates correspond to the local maxima of the autocorrelation function of the low-pass filtered input LP residual signal. The second stage is applied only to the voiced speech and the transition speech for voiced/transition speech classification and the final pitch determination. In this stage, the AbS parameter estimation procedure illustrated by Figure 1 is employed. For each time/spectral pitch candidate, signal modification is performed on the original LP residual signal. The final pitch decision is made based on the closed-loop error signal for each pitch candidate. At the last stage, a pitch refinment and harmonic bandwidth estimation procedure similar to [6] is performed on subframes that are declared as voiced. The signal modification procedure we use is similar to that used in the EVRC coder [9].

The closed-loop information from the time scale signal modification is also used to improve the voiced/transition classification. This approach is based on the assumption that if a voiced subframe is hypothesized, we should be able to find a reasonable pitch value for that subframe that 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 this closed-loop pitch estimation algorithm significantly reduces gross pitch errors when compared to a conventional time domain pitch estimator based on the autocorrelation function. For example, the percentage of pitch outliers that have normalized pitch error larger than 10% is reduced from 7.3% to 1.2%. We also conducted an A/B subjective test to compare the synthesized speech using manually estimated pitch and classification with the synthesized speech using the proposed AbS pitch estimation/classification scheme. In the test, 16 speech sentence pairs were played for 11 listeners. The test indicates a preference of 38.6% for the use of manually estimated parameters versus 38.1% for the use of proposed AbS parameter estimation algorithm, with 23.3% for no preference. This result indicates that there is practically no degradation in the proposed pitch and class estimation when compared to the ideal manual segmentation.

3. TRANSITION SPEECH CODING

Though the harmonic model is well-suited for the reconstruction of voiced and unvoiced speech signals, it is ineffective for representing the transition speech signals such as voicing onsets, plosives and non-periodic pulses. The AbS-MHC coder employs a novel frequency domain transition model which replaces the conventional harmonic model used for voiced/unvoiced speech signals. In this new transition model, the LP residual signal is represented by the following generalized sinusoidal model:

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

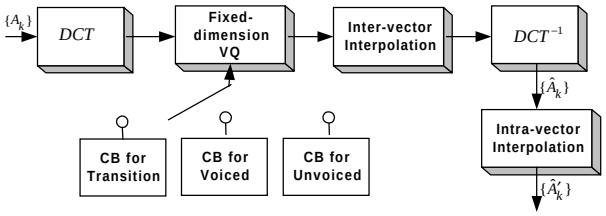

Figure 2: Spectral magnitude vector quantization

with the phase θ_k given by

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

where N is the frame or subframe size in samples, $\{n_j\}$ 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 (10 ms in our coder) so that it is reasonable for all the pulses to use the same spectral envelope parameters with different gains $\{g_i\}$.

In this model, pulse occurrence information, which is the most important information in a transition frame, is represented by parameters $\{n_j\}$. The dispersion phase $\{\varphi_k\}$ and spectral magnitude $\{A_k\}$ represent together the pulse shapes. Experimental evidence shows that coarse quantizations of A_k and φ_k are perceptually acceptable. In our AbS-MHC coder, the magnitude vector for transition frames is obtained by uniformly sampling the LP residual spectrum at 100 Hz intervals. The magnitude vector is then quantized and linearly interpolated in the DCT transform domain. The other parameters are estimated by minimizing the weighted mean squared error:

$$E = \sum_{m=0}^{N-1} \left(S_w(n) - h_w(n) * \sum_j g_j \sum_k A_k \cos\theta_k(n, n_j)\right)^2 \tag{5}$$

where $S_w(n)$ is the weighted original transition speech signal, $h_w(n)$ is the impulse response of the weighted LP synthesis filter, and $*$ stands for the convolution operation. The components of the dispersion phase vector are set equal to a fixed value that is found by closed-loop quantization. The pulse position parameters $\{n_j\}$ are found using a sequential search whereby for each value n_j, the optimal value of the corresponding gain g_j is computed and used in the search criterion (5). To reduce the encoding rate for $\{n_j\}$, a constrained codebook for shift parameters is used. For each subframe, all the positions are divided into several grid tracks and one pulse is found on each track. Once $\{A_k\}$, $\{n_j\}$, and φ are obtained, the optimal gain vector $\{g_j\}$ is computed and quantized by mean removed VQ.

4. WNSTVQ BASED SPECTRAL MAGNITUDES QUANTIZATION

Quantization of the variable-dimension harmonic magnitude vector is one of the most challenging problems in harmonic speech coding. Our quantization scheme uses DCT-II transform based WNSTVQ method [10] in conjunction with intra/inter-vector interpolations. A simplified block diagram for this approach is shown in Figure 2.

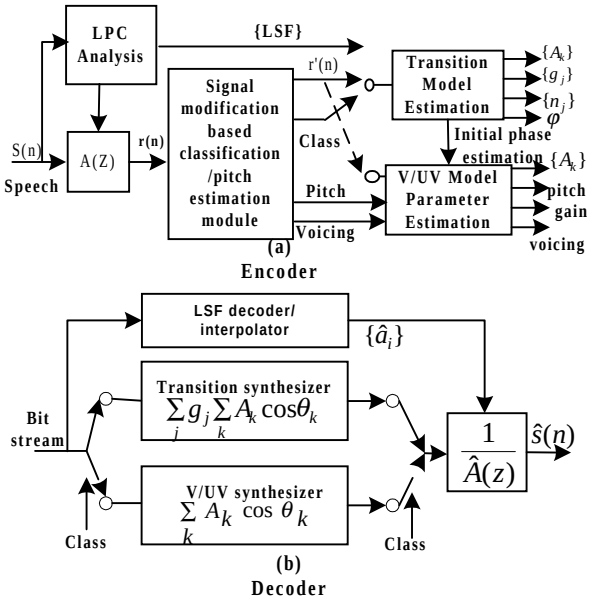

Figure 3: Simplified block diagram of AbS-MHC speech codec

The harmonic magnitude vector $\{A_k\}$ is obtained by sampling the LP residual magnitude spectrum at pitch harmonics. For components above the voicing frequency, the average values around the pitch harmonics are used. This variable-dimension vector is DCT transformed to the fixed-dimension vector domain and quantized by the WNSTVQ method [10]. Voiced speech, unvoiced speech and transition speech have quite different spectral characteristics, hence we designed different spectral codebooks for each class. Experimental evidence shows that, for our coder, quantizing spectral vectors every frame and reconstructing the vector every subframe by linear vector interpolation in the fixed-dimension domain provided better perceptual quality than quantizing spectral vectors every subframe using half of the frame bit rate.

For voiced speech, the spectrum is sampled at pitch harmonics over the whole frequency band. In the synthesis, the nonharmonic components above the voicing frequency are actually reconstructed at 100 Hz intervals instead of pitch intervals. To recover the nonharmonic components, the linear intra-vector interpolation is applied to the reconstructed variable-dimension vector $\{\hat{A}_k\}$.

5. DESIGN OF THE 4 KB/S ABS-MHC CODER

We designed a 4 kb/s AbS-MHC coder to test in a real coding environment the proposed new techniques described in Section 2, 3, and 4. The simplified block diagram of the AbS-MHC codec is shown in Figure 3. 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. The AbS parameter estimation module described in Section 2 is used to estimate the class, the pitch, and the voicing parameters. The encoder is set to a particular encoding mode according to the classification. For voiced/unvoiced frames, the speech model (1) is used and harmonic magnitudes $\{A_k\}$ are quantized by the procedure described in Section 4. For transition frames, the transition model (3) and the parameter quantization scheme described in Section 3 are applied. According to the received classification 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.

For operation at 4 kb/s, a frame length of 20 ms (160 samples at 8 kHz sampling rate) is used. Each frame is divided into 2 subframes. The bit allocation for voiced/unvoiced modes is given in Table 1.

Parameters	1st subframe	2nd subframe	frame
LSFs		4×6	24
class			1
pitch	5	7	12
harmonic magnitudes	0	4×7	28
voicing	0	6	6
gain	0	7	7
total			78

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

The bit allocation for transition mode coding at 4 kb/s is given in Table 2.

Parameters	1st subframe	2nd subframe	frame
LSFs		3×6	18
class			1
shifts $\{n_j\}$	5+5+4	5+5+4	28
pulse signs	3	3	6
pulse gains	6	6	12
excitation spectral $\{A_k\}$	0	7+6	13
dispersion phase	1	1	2
total			80

Table 2: Bit allocation for the transition mode at 4 kb/s

6. SUBJECTIVE TEST RESULTS

To evaluate the performance of the 4 kb/s AbS-MHC coder, we first ran an informal A/B (pairwise) listening tests. Eleven listeners compared the 4 kb/s AbS-MHC coder with the G.723.1 coder operating at 6.3 kb/s. Sixteen sentences spoken by 8 male and 8 female speakers were used. All these sentences were filtered by the modified IRS filter. Test results are shown in Table 3. We also conducted an absolute category rating (ACR) subjective quality test to obtain the Mean Opinion Score (MOS) for the 4 kb/s AbS-MHC coder. For reference, the 4 kb/s MPEG-4 standard coder, G.723.1 coder operating at both 5.3 kb/s and 6.3 kb/s, and 32 kb/s G.726 ADPCM coder were included. The speech material for the test consisted of 16 sentence pairs, 8 from female talkers and 8 from male talkers. They were all filtered by the modified IRS filter and normalized to -26 dB. Fourteen listeners participated in the test, and the test results are summarized in Table 4.

The subjective results show that the 4 kb/s AbS-MHC coder outperforms the 4 kb/s MPEG-4 standard and the 5.3 kb/s G.723.1 CELP coder, and it produces speech quality very similar to the 6.3 kb/s G.723.1 coder.

	Pref. G.723.1 (6.3 kb/s)	Pref. AbS-MHC (4 kb/s)	No pref.
Overall	36.7%	43.8%	19.5%
Female	32.8%	40.6%	26.6%
Male	40.6%	46.9%	12.5%

Table 3: A/B test results

coder	MOS overall	MOS female	MOS male
4 kb/s MPEG-4	2.57	2.39	2.74
5.3 kb/s G.723.1	2.99	2.92	3.05
4 kb/s AbS-MHC	3.39	3.43	3.35
6.3 kb/s G.723.1	3.42	3.36	3.48
32 kb/s G.726	3.58	3.47	3.70

Table 4: MOS test results

7. 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. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. on Speech and Audio Processing*, pp. 242–250, July 1995.

[6] E. Shlomot, V. Cuperman, and A. Gersho, "Combined harmonic and waveform coding of speech at low bit rates," in *Proceedings of ICASSP*, pp. 585–588, 1998.

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

[8] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proceedings of ICASSP*, pp. 249–252, 1990.

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

[10] 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.