COMBINED HARMONIC AND WAVEFORM CODING OF SPEECH AT LOW BIT RATES

Eyal Shlomot, Vladimir Cuperman and Allen Gersho

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

ABSTRACT

In this paper we present a new approach for speech coding, which combines frequency-domain harmonic coding for periodic and "noise like" unvoiced segments of speech with a time-domain waveform coder for transition signals. This hybrid coder requires special handling of the boundary between voiced and transition segments. We outline the details of a 4 kbps hybrid coder and present subjective quality test results of this coder.

1. INTRODUCTION

Modern CELP coders, using analysis-by-synthesis within the excitation-filter framework, are able to produce high quality speech at rates down to 6 kbps, but are incapable of delivering high quality speech at lower bit rates. Parametric vocoders which make use of the harmonic structure of the spectrum during voiced speech and the "noise like" characteristics of unvoiced speech, can compress speech with high intelligibility and reasonable quality at a bit rate as low as 2.4 kbps [1].

Waveform coders fail at low bit rates since they try to represent the perceptually unimportant waveform shape. But why do the parametric "harmonic coders", using a harmonic model for voiced speech and a noise model for unvoiced speech, fail to deliver higher quality speech? Examining Fig. 1 we can see vowel segments which have strong periodic characteristics and fricative segments which have a stationary "noise like" characteristic. But we can also clearly observe transition segments, which are neither periodic nor "noise like". These segment, such as onsets, plosives and non-periodic glottal pulses, consist of local time events which cannot be represented by the periodic or the noise models (or even a combination of both). Transition segments are only a small percentage of the speech signal, but convey a lot of information and their faithful reproduction seems to be important for high quality speech.

To overcome the harmonic coders limitations we added a third model for the representation of transition segments, employing a time-domain waveform coder to capture the location and structure of the local time events. We call this concept "hybrid coding", marking the integration of a time-domain waveform coder with a frequency-domain parametric coder [2].

The addition of a time-domain module for the local time events seems like a natural choice, but its interoperability with a frequency-domain coder creates a synchronization problem. At low bit rate harmonic coding, the linear

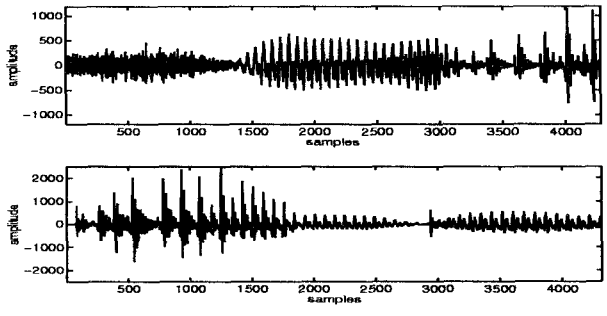

Figure 1: Samples of a Speech Waveform

phase (i.e., a time shift) information is not transmitted and therefore the reconstructed harmonic signal is not synchronized with the original one. On the other hand, a waveform coder generates a signal which is time aligned with its target signal. Therefore, signal continuity is not preserved when switching from one model to the other. In this work we present novel phase synchronization modules which provide signal continuity without transmitting additional information.

We designed a 4 kbps coder, demonstrating the hybrid coding concept. Our harmonic and "noise like" building blocks, although conceptually similar to other low rate parametric coders, include several new and novel concepts. For example, novel perceptually-weighted quantization scheme, within the general linear dimension conversion method, was used for harmonic spectral quantization. Other unique features include a multi-pulse excitation and a closed-loop analysis-by-synthesis search algorithm for time-domain waveform coding of transition segments, and a neural-network classifier, trained by a large training set, to obtain the speech class.

In section 2 we describe the general concept of the hybrid coding scheme, while in section 3 we outline the structure of the 4 kbps hybrid coder. In section 4 we present the results of a subjective listening test for the 4 kbps hybrid coder.

2. GENERAL DESCRIPTION OF HYBRID CODING

A schematic diagram of a hybrid encoder is presented in Fig. 2, and a schematic diagram of a decoder in Fig. 3. A Linear Prediction (LP) module is used to obtain the

0-7803-4428-6/98 $10.00 © 1998 IEEE

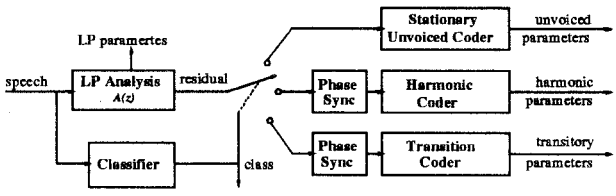

Figure 2: A Schematic Diagram of an Hybrid Encoder

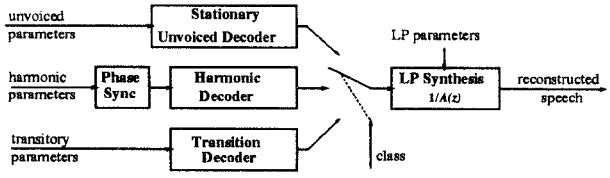

Figure 3: A Schematic Diagram of an Hybrid Decoder

residual signal, which is the target signal for the hybrid encoder. The speech classifier assigns a class decision (harmonic, transition or stationary unvoiced) for each frame, which is used to control a switch between the three possible coders. The LP parameters, the class decision and the quantized parameters from the appropriate coder are sent for each frame. The excitation signal is decoded according to the transmitted class and passed through a synthesis LP filter to generate the reconstructed speech signal.

Any waveform coding scheme, representing the local time events of the transition frames, can be used in the hybrid coder.

In the following, we review a general parametric approach of frequency-domain harmonic coding and describe the novel phase synchronization methods which enable the integration of such harmonic coding schemes with a waveform time-domain coder.

2.1. The Harmonic Coder

Harmonic speech segments can be approximated using only samples of the spectral harmonic peaks, which hold most of the signal information. Each spectral peak, indexed by k, represents an oscillator $o_k(t)$, given by

$$o_k(t) = a_k \cos(k\theta_0 + k2\pi f_p t + \psi_k) \tag{1}$$

where f_p is the pitch frequency (assuming the pitch frequency does not change during the spectral analysis frame), θ_0 is the constant linear phase and $2\pi f_p t$ is the harmonic phase. The linear and harmonic phases of all oscillators are related by the index k and are completely defined by θ_0 and f_p. The "dispersion" phase, ψ_k, is unique for each peak and dictates the local pulse structure.

The human auditory system is insensitive to the linear phase, as long as phase continuity is preserved. It can also tolerates an inaccurate or completely absent dispersion phase. These perceptual properties are important factors in the success of the harmonic models for low bit rate coding, where a synthetic phase replaces the measured one. The synthetic phase is generated by disregarding the dispersion phase, but taking into account the time dependency of the pitch frequency. The k^{th} oscillator is given by

$$o_k(t) = a_k \cos[k\theta(t)] \tag{2}$$

with the continuous phase evolution

$$\theta(t) = \theta_0 + 2\pi \int_{t_0}^{t} f_p(\tau) d\tau. \tag{3}$$

When the discrete measurements of the pitch frequency are linearly interpolated, the phase evolution becomes

$$\theta(t) = \theta_0 + 2\pi \left[f_p(n-1)t + \frac{1}{2T}[f_p(n) - f_p(n-1)]t^2 \right], \tag{4}$$

where $f_p(n-1)$ and $f_p(n)$ are the pitch frequencies of the previous and the current frame respectively, and T is the pitch sampling interval. Phase continuity is preserved if θ_0 of the next frame is taken as the phase evolution value at the end of the current frame.

The quantization of the variable dimension vector of spectral peaks magnitudes, $\{a_k\}$, is discussed in section 3.3. The decoder generates the excitation for each frame using a sum of all the harmonic oscillators, combining the quantized spectral magnitudes and the synthetic phase:

$$ex(t) = \sum_k a_k^q \cos[k\theta(t)]. \tag{5}$$

Overlap-and-add can be used to help in signal smoothing between frames.

2.2. The Stationary Unvoiced Coder and Mixed Signals Coding

The complex structure of stationary unvoiced speech segments have no perceptual importance, hence they can be adequately represented using white noise modulated by the energy contour and the spectral envelope. However, some portions of the speech (sometimes called "mixed signals") are neither completely voiced nor unvoiced. We represent mixed signals using two spectral bands, a low harmonic band and a high non-harmonic band. We call the width of the lower harmonic band the "harmonic bandwidth", which is one of the transmitted parameters. The harmonic bandwidth can be zero, indicating stationary unvoiced speech, or its value can specify any other mixture of a lower harmonic spectrum and a higher non-harmonic spectrum.

The oscillators of the low harmonic band use the synthetic phase given by Eq. (4) and the oscillators of the high non-harmonic band use a uniformly distributed random phase.

2.3. Switching from Transition Segment to Harmonic Segment

The reconstructed transition signal is time-aligned to the original reference frame, but since the initial linear phase of the harmonic segment, θ_0, is not transmitted by a typical low bit rate coder, the following reconstructed harmonic segment is not synchronized with the original one. Therefore, when switching from a transition frame to a harmonic frame, signal continuity at the frame boundary is not preserved.

An initial linear phase estimation, obtained by maximizing the correlation of the shifted reconstructed harmonic excitation frame with the reconstructed transition excitation frame, is used for synchronization during switching. Since the correlation is computed between the coded segments, this estimate can be performed by the decoder, without transmitting any additional information. The estimated linear phase is propagated through the harmonic segments by the phase evolution formula (Eq. (4)).

2.4. Switching from Harmonic Segment to Transition Segment

On harmonic segments, linear phase deviation can occur between the original signal and the synthesized one. This is the result of a possibly inaccurate onset synchronization, pitch estimation and quantization errors, as well as the approximation and the discrete nature of the phase evolution formula. Linear phase deviation means a loss of synchronicity between the reconstructed harmonic frame and the original harmonic frame. Since the following transition frame is time aligned with the original one, signal continuity is lost at the frame boundary.

To synchronize the reconstructed transition frame with the preceding harmonic one, the drift between the original signal and the reconstructed one is measured by the encoder, and the transition frame is extracted with the corresponding shift. The same shift is used for all the frames of the transition segment.

2.5. Switching to and from Stationary Unvoiced Segment

The phase information is not essential for reproducing the stationary unvoiced segments. Therefore no phase synchronization is required when switching to or from such segments. Moreover, a calculated phase correction term, which is carried over all frames of a harmonic speech segment or a transition speech segment, can be reset when a stationary unvoiced segment is encountered.

3. THE HYBRID 4 KBPS CODER

We designed a 4 kbps coder to demonstrate the hybrid concept. The switching schemes presented in sections 2.3, 2.4 and 2.5 were used for this coder. The following sections outline some details of the structure of the 4 kbps hybrid coder.

3.1. General Structure and Linear Prediction Analysis

The 4 kbps coder operates on a telephone bandwidth speech sampled at the rate of 8 kHz. The DC component and low-frequency rumble are removed by an eighth-order IIR high-pass filter with the cutoff frequency of 50 Hz. The LP analysis, performed every 20 ms frame, is very similar to the one suggested for the ITU-T Recommendation G.729 [3]. It utilizes an asymmetric window with 5 ms lookahead, bandwidth expansion and high frequency compensation performed on the autocorrelation function. The LP coefficients are converted to the LSFs representation and quantized by a predictive multi-stage quantizer, using 18 bits in two stages of 9 bits each. The quantized LSFs are interpolated every 5 ms and then converted back to prediction coefficients which are used by the inverse filter to generate the residual signal. Unquantized LSFs are also interpolated and converted to unquantized prediction coefficients, to generate the perceptually weighted speech used as a reference signal for transition frames.

3.2. Classification, Pitch and Harmonic Bandwidth Estimation and Quantization

The classifier module in Fig. 2 serves as a pitch and harmonic bandwidth estimator as well. For each 10 ms subframe we compute a vector of classification parameters. The vector is formed by the concatenation of three sets, representing the signal for the past, current and future subframes, and includes speech energy, spectral tilt, rate of zero-crossing, residual peakiness, residual harmonic matching SNRs and pitch deviation measures.

The classification parameters, including the network decision from the previous frame, are fed into a three layers, fully connected feed-forward neural network. The winning output from the three neurons of the output layer specifies the class, but some manually tuned hysteresis was added to avoid classification "jitter".

For harmonic frames, the residual harmonic matching SNRs are also used to determined the pitch frequency and the harmonic bandwidth.

The pitch frequency is quantized in the range of 60 Hz to 400 Hz, using a 7 bits uniform quantizer. The harmonic bandwidth is quantized using only 3 bits. The quantized harmonic bandwidth value of zero is used to indicate a stationary unvoiced segment, hence the class information requires only one bit, indicating transition or non-transition subframe.

3.3. Quantization of Harmonic Spectral Envelope

For each 10 ms subframes, a spectral representation of the residual signal is obtained using a Hamming window centered at the middle of the subframe and a 512 point DFT. The harmonic peaks of the residual magnitude spectrum, within the harmonic bandwidth, are sampled at the multiples of the pitch frequency. At frequencies above the harmonic bandwidth the spectrum is represented by an average of the samples around the multiples of the pitch frequency. The sampling (and averaging) procedure generates an M dimensional vector, where the variable dimension M is inversely proportional to the pitch frequency.

Variable dimension vector quantization can be achieved by converting the variable dimension vector to a fixed dimension vector which is then quantized. Dimension conversion methods can be either linear or nonlinear. By linearity we mean that the fixed dimension vector is a linear (pitch dependent) function of the variable dimension vector. An example of a nonlinear method is the DAP [4] algorithm, and an example of a linear scheme is the VDVQ method [5], where spectral samples are mapped into spectral bins. The general form of linear dimension conversion was presented in [6] under the name Non-Square Transform (NST), where a fixed dimension vector y is generated from the variable

dimension vector x by multiplying x with a non-square matrix B of dimension $N \times M$. The matrix B is one of a family of matrices, since the dimension of B depends on M, which in turn depends on the pitch frequency. An exact or approximate version of x (depending on the properties of B) can be recovered by $x = Ay$.

Here we address the issue of Weighted Mean Square Error (WMSE) minimization using the NST. A "closed-loop" error minimization on the residual spectrum was suggested in [7] where a weighted distance, combining the spectral magnitude of the LP synthesis filter and a perceptual weighting filter, was used between the spectral vector x and the quantized vector x_q. The WMSE, ϵ, is given by

$$\epsilon = (x - x_q)^T W (x - x_q), \tag{6}$$

where W is a diagonal matrix given by

$$W_{kk} = \left\| \frac{A(z/\gamma_1)}{A(z)A(z/\gamma_2)} \right\|^2_{z=exp\left(j\frac{2\pi k f_p}{F_s}\right)}. \tag{7}$$

Since the quantization is performed on the fixed dimension vector y, and since $x = Ay$, equation 6 takes the form of

$$\epsilon = (Ay - Ay_q)^T W (Ay - Ay_q) = (y - y_q)^T A^T W A (y - y_q). \tag{8}$$

From practical computation considerations, the transform matrices pair A and B should be selected such that $A^T W A$ is a diagonal matrix. It can be easily shown that $A^T W A$ is a diagonal matrix for the VDVQ method as well as for the simple method of dimension conversion using zero-padding.

To select between the VDVQ and the zero-padding methods we run a test, measuring the average WMSE on a large database. For both methods, in order to capture the varying characteristics of the spectral vector at different pitch values, the pitch frequency range was divided into 6 zones, and a unique codebook was designed for each zone. The zero-padding method performed slightly better than the VDVQ method, and was used for the 4 kbps coder. The codebooks for the zero-padded harmonic spectral envelope use 14 bits in a two-stage structure and 6 bits are used for the gain.

3.4. Coding of Transition Signal

For time-domain excitation coding of transition frames we use five signed pulses and one gain for each 10 ms subframe. The pulse locations are confined to a grid, and the sign for each pulse is set according to the instantaneous sign of the residual signal. The optimal pulse locations are determined by a full search analysis-by-synthesis scheme and 19 bits are used to describe the pulse locations, 5 bits for the signs and 6 bits for the gain.

4. SUBJECTIVE TEST RESULTS

We conducted an absolute category rating (ACR) subjective quality test to obtain the Mean Opinion Score (MOS) for the 4 kbps hybrid coder. For reference, two CELP-based standard coders were included in the test: the Federal Standard 1016 at the rate of 4.8 kbps and the ITU-T Recommendation G.723.1 at the rate of 5.3 kbps. The speech material for the test consists of 16 sentence pairs, 8 from female talkers and 8 from male talkers. To simulate the telephone bandwidth condition, the speech material was filtered by the modified IRS filter and the test was conducted using a telephone handset. Ten non-expert listeners participated in the test, and the results are summarized in Table 1.

coder	MOS female	MOS male	MOS total
4.8 kpbs FS-1016	2.65	3.21	2.93
4.0 kbps hybrid coder	3.23	3.55	3.39
5.3 kbps G.723.1	3.21	3.68	3.45

Table 1: MOS results

The MOS test results show that our 4 kbps coder performs much better than the 4.8 kbps Federal Standard 1016 and that its quality is close to the ITU-T Recommendation G.723.1 at the rate of 5.3 kbps. These results suggest that the hybrid coding approach introduced here has the potential of competing favorable with CELP coding at rates of 4 kbps and below.

The authors wish to thank Ashish D. Aggarwal and Cagri Etemoglu for conducting the MOS test.

5. REFERENCES

[1] W. Kleijn and K. Paliwal, eds., *Speech Coding and Synthesis.* Amsterdam: Elsevier Science Publishers, 1995.

[2] E. Shlomot, V. Cuperman, and A. Gersho, "Hybrid coding of speech at 4 kbps," in *Proceedings of the IEEE Speech Coding Workshop*, (Pocano Manor, Pennsylvania, USA), pp. 37–38, 1997.

[3] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Description of the proposed ITU-T 8 kb/s speech coding standard," in *Proceedings of the IEEE Speech Coding Workshop*, (Annapolis, Maryland, USA), pp. 3–4, 1995.

[4] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 39, pp. 441–423, Feb. 1991.

[5] A. Das, A. Rao, and A. Gersho, "Variable-dimension vector quantization of speech spectra for low-rate vocoders," in *Proceedings of the Data Compression Conference*, pp. 421–429, 1994.

[6] P. Lupini and V. Cuperman, "Non-square transform vector quantization for low-rate speech coding," in *Proceedings of the IEEE Speech Coding Workshop*, (Annapolis, Maryland, USA), pp. 87–89, 1995.

[7] M. Nishiguchi and J. Matsumoto, "Harmonic and noise coding of LPC residuals with classified vector quantization," in *Proceedings of the IEEE International Conference on Acoustics, Speech & Signal Processing*, pp. 484–487, 1995.