HYBRID CODING OF SPEECH AT 4 KBPS

Eyal Shlomot, Vladimir Cuperman, and Allen Gersho

Department of Electrical and Computer Engineering,
University of California, Santa Barbara, CA 93106
gersho@ece.ucsb.edu

ABSTRACT

In this paper we present a novel scheme for hybrid coding of speech signals. This hybrid codec utilizes the excitation/filter model used extensively for speech coding. Similar to other modem vocoders, voiced speech is represented by a frequency domain harmonic model and unvoiced speech by a "noise-like" excitation. However, an analysis-by-synthesis time domain scheme is employed for the transitory portions of the speech signal which cannot be adequately represented by either model. Switching between the time domain and the frequency domain models requires careful handling of the reconstructed linear phase. The structure of a 4 kbps speech codec, based on the hybrid model, is outlined in this paper. The new codec shows promise of achieving toll quality at 4 kbps.

1. INTRODUCTION

Waveform and parametric coding are the two distinct paradigms for speech compression. Waveform coding, such as PCM, ADPCM and CELP, are based on some matching between the reconstructed waveform and the original. Parametric speech compression, on the other hand, is achieved by coding parameters of a speech generation model without attempting to match the reconstructed waveform to the original.

Traditionally, speech segments have been classified as either voiced or unvoiced for the purpose of modeling and coding. Harmonic models in the frequency-domain were extensively investigated for voiced speech and "noise-like" models were used to represent the unvoiced portions of the speech [1-6]. While the harmonic models for voiced speech and the "noise-like" models for fricatives are well justified and perform adequately, these models are unsuitable for onsets, plosives and non-periodic glottal excitation. Listening tests demonstrate a degradation in the speech quality if either harmonic or "noise-like" models are used for these parts of the speech signal.

In this paper, we present a novel hybrid model for speech coding. The hybrid scheme is based on Linear Prediction (LP) and parametric modeling of the excitation signal. We classify the speech signal into voiced, unvoiced or transitory portions, and generate a suitable type of excitation for each class. Frequency domain harmonic excitation is used for voiced speech, "noise-like" excitation represents the unvoiced segments, while multipulse excitation is employed for the transitory portions. The combination of the multipulse time domain scheme with the frequency domain harmonic coding necessitates the development of phase synchronization modules for the handling of the reconstructed linear phase. Following the hybrid concept, we developed a 4 kbps hybrid codec and describe some of its details in this paper.

2. HYBRID CODING OF SPEECH

In the hybrid encoder, the LP parameters are estimated using well-known methods and a classifier/pitch/voicing module makes the class decision and estimates the pitch frequency and the level of voicing (the size of the harmonic portion of the spectrum). Frequency domain harmonic modeling is performed, starting with a windowed DFT and obtaining sampled and quantized harmonics, to represent the excitation for voiced speech. A dense uniform-interval frequency magnitude sampling, combined with a random phase in the decoder, yield a good reproduction of unvoiced fricative sounds. Mixed voiced-unvoiced segments are modeled by a combination of these models, controlled by the level of voicing. While voiced and unvoiced speech share a frequency domain model, a time domain analysis-by-synthesis waveform matching model is used for the transitory portions.

The linear phase of the harmonic model of a voiced frame must be synchronized with the reconstructed signal of the transitory frame which precedes it. Similarly, a transitory frame which follows a voiced frame must be properly aligned. Since no phase information is sent from the encoder to the decoder, the linear phase synchronization and the alignment factor must be resolved by the decoder and the encoder without side information.

3. THE 4 KBPS CODEC

The 4 kbps hybrid codec uses a frame size of 20 ms with a 25 ms look-ahead. The LP analysis is performed using a non-symmetric window, and the LSFs are quantized using predictive multistage VQ. Classification, pitch and voicing are obtained for each 10 ms sub-frame, and either voiced, unvoiced or transitory coding is performed on each sub-frame unit.

3.1. The Classifier/Pitch/Voicing Module

Robust classification, pitch detection and voicing is essential for the hybrid codec. We use frequency domain matching algorithm, similar to [3], but operating on the LP residual signal. The DFT magnitude of the residual signal is matched to a set of frequency domain synthetic comb representations of the window, build of 2, 4 and 6 lobes and guided by the SEEVOC envelope. The frequency which maximizes the SNR between the residual magnitude and the synthetic combs, is chosen as the initial pitch frequency.

0-7803-4073-6/97/$10.00©1997IEEE.

The set of classification parameters consists of the frequency matching maximal SNR, energy, zero-crossing rate, spectral flatness and pitch deviation. These parameters, derived from the previous, current and the next frame, are used by a feed-forward neural network to obtain the classification decision.

The frequency domain matching algorithm is also used to determine the level of voicing. The classification decision, the pitch frequency and the level of voicing are smoothed, to avoid rapid changes which can generate audible artifacts.

3.2. The Voiced and Unvoiced Models

The harmonic model for voiced speech is based on the assumption that the perceptually important information resides essentially at the harmonic samples of the pitch frequency. These samples are complex valued, providing both magnitude and phase information. The phase information consists of three terms; the linear phase, the harmonic phase and the dispersion phase. The linear phase component is simply the time shift of the signal, the harmonic phase is the integral of the pitch frequency and the dispersion phase governs the structure of the pitch pulse and is related to the structure of the glottal excitation.

For low bit-rate sinusoidal coding, the linear and dispersion terms of the phase are usually discarded, and the harmonic phase is reconstructed solely as an approximated integral of the pitch frequency, with an arbitrary linear phase shift. In our hybrid codec, a time domain transitory model is used for the onset which usually precedes the voiced portions of the speech. An arbitrary linear phase applied to the harmonic model for the voiced frame will create a discontinuity of the signal on the frames boundary. Hence, the decoder linear phase synchronization module estimates the linear phase component, using the correlation between the transitory frame and the voiced frame. The estimated linear phase is used as the initial phase for the harmonic phase generator.

The sampled magnitudes at the harmonics of the pitch frequency form a variable dimension vector. The quantization of such a variable dimension vector is an interesting problem which was addressed extensively elsewhere and has many possible solutions [4][5]. We use the concept of dimension conversion into a fixed dimension vector, and employ a perceptually weighted multistage predictive Vector Quantization (VQ) to represent this vector [6].

The quantized spectral peaks, combined with the synthesized harmonic phase and the estimated linear phase, are used to generate the harmonic signal for voiced speech.

A similar approach is used to generate the "noise-like" signal. The frequency magnitude is sampled on equal intervals, and a uniformly distributed random phase is applied to each frequency component. This unified approach for the representation of voiced and unvoiced speech enables easy voicing control by selecting a frequency band in which the harmonic phases are randomized.

3.3. The Transitory Model

We use a multipulse excitation for the transitory portions of the speech signal, such as onsets, plosives and non-periodic glottal excitation. The multipulse excitation is comprised of a few pulses located on a grid. A gain term can be assigned to each pulse, or each pulse can be signed and multiplied by a common frame gain. An analysis-by-synthesis pruned-tree search method is employed to find the pulse locations and the gains. Scalar predictive quantization are used for the common frame gain, while multiple gains are quantized using VQ.

A transitory frame which follows a voiced frame must be aligned with the preceding reconstructed signal to avoid discontinuities. The alignment is solved by the encoder synchronization module which uses the past reconstructed signal as a "phase reference" for the encoding of the current frame.

4. RESULTS

The proposed codec was tested using informal subjective MOS evaluation. The unquantized model (unquantized pitch, voicing, gains, LPC coefficients, and harmonic magnitudes) achieved quality comparable to the low-rate G.723.1 codec.

As an initial feasibility test, a crude quantization of parameters was performed with the bit allocation given in Table 1 and the resulting 4 kbps codec had a quality comparable to the IMBE 4.1 kbps standard. We expect that optimization of the bit allocation and quantization techniques (currently in progress) will yield a performance that is very close to toll quality.

Harmonic Parameters	Bits	Transitory Parameters	Bits
LSFs	16	LSFs	16
Class	1x2	Class	1x2
Pitch	5x2	Pulse Locations	16x2
Voicing	3x2	Pulse Gains	15x2
Gain	5x2		
Harmonic Peaks	18x2		
Total	80	Total	80

TABLE 1: Bit Allocation for the 4 kbps codec

References

[1] L.B. Almeida and J.M. Tribolet, "Non-stationary spectral modeling of voiced speech", *IEEE Trans. on ASSP*, Vol. 31, No. 3, pp. 664-678, June 1983.

[2] Y. Shoham, "High-quality speech coding at 2.4 and 4.0 kbps based on time-frequency interpolation", *Proc. of IEEE ICASSP*, pp. II-167 - II-170, 1993.

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

[4] A. Das, A.V. Rao, and A. Gersho, "Variable dimension vector quantization", *IEEE Sig. Proc. Let.*, pp. 200-202, July 1996.

[5] V. Cuperman, P. Lupini and B. Bhattacharya, "Spectral excitation coding of speech at 2.4 kb/s", *Proc. of IEEE ICASSP*, pp. 496-499, 1995.

[6] M. Nishiguchi and J. Matsumoto, "Harmonic and noise coding of LPC residuals with classified vector quantization", *Proc. of IEEE ICASSP*, pp. 484-487, 1995.