A VARIABLE-RATE MULTIMODAL SPEECH CODER WITH GAIN-MATCHED ANALYSIS-BY-SYNTHESIS

Erdal Paksoy, Alan McCree, and Vishu Viswanathan

Corporate Research, Texas Instruments, Dallas, TX

ABSTRACT

In general, a variable rate coder can obtain the same speech quality as a fixed rate coder, while reducing the average bit rate. We have developed a variable-rate multimodal speech coder with an average bit rate of 3 kb/s for a speech activity factor of 80% and quality comparable to the GSM full rate coder. The coder has four coding modes and uses a robust classification method involving the pitch gain, zero crossings, and a peakiness measure. Also the coder employs a novel gain-matched analysis-by-synthesis technique for very low rate coding of unvoiced frames and an improved noise-level-dependent postfilter. This paper describes the details of our algorithm and presents the results from subjective listening tests.

1. INTRODUCTION

Speech coding is an important part of digital voice communication and storage systems. By reducing the number of bits required to adequately represent the speech signal, a speech coder increases the capacity in such applications. Variable rate coding [1] is well suited for storage systems since these are not subject to the fixed transmission rate constraint often imposed by communication systems.

We have developed a high-quality multimodal variable rate speech coder intended for digital telephone answering machine applications. This work was based on a low rate version of the code excited linear prediction (CELP) coder we developed for the GSM enhanced full rate standardization activity [2]. The coder operates at an average bit rate of 3 kb/s and classifies speech into one of four coding modes, which roughly correspond to (1) background noise, (2) voiced speech, (3) steady-state voiced speech, and (4) unvoiced speech. Different bit allocations and coding strategies are used in each mode. We use several new techniques: a peakiness measure for use in the mode decision, a new voice activity detector (VAD), reduced fluctuation quantization of gain and linear predictive coding (LPC) parameters for noise frames, and gain-matched analysis-by-synthesis coding for unvoiced frames.

The paper is organized as follows. In Section 2, we present an overview of the coding algorithm. In Section 3, we describe in detail the mode decision mechanism. In Section 4, we describe the various coder components. Section 5 summarizes the bit allocation for the different modes. Section 6 describes the listening tests we conducted to compare the coder with several reference coders.

2. OVERVIEW

The coder has a frame size of 40 ms, with 4 subframes of 10 ms each. The classification of the signal into one of four modes is done in open-loop fashion. Below we present an overview of these four modes along with their coding strategies.

1. **The Noise Mode** is designed to code portions of the input where no speech is present. The excitation is obtained by simply gain-scaling the output of a random noise generator. This is the mode with the lowest bit-rate since only the gain and LPC parameters are transmitted.
2. **The Unvoiced Mode** is designed for frames characterized by an absence of pitch periodicity. This mode is primarily used for unvoiced fricatives such as /s/, /sh/, /f/, /th/. Therefore, the adaptive codebook is not used. The LPC parameters are coded at a lower rate, and the excitation is chosen from a sparse Gaussian codebook.
3. **The Voiced/Transient Mode** has the highest bit rate and is used for (non-steady-state) voiced speech, for which the pitch information as well as the spectral envelope need to be coded with precision. Hence we use an adaptive codebook with fractional sample resolution and a fixed ternary pulse (+1, -1, 0) excitation codebook [2]. This mode is also appropriate for transitions between voiced and unvoiced speech and for unvoiced plosives such as /p/, /t/, /k/, which contain short bursts of energy followed by short silences and which are thus well represented with the pulse codebook.
4. **The Steady-State Voiced Mode** is very similar to the Voiced/Transient Mode, but is intended to encode voiced frames during which the LPC and pitch information vary slowly. Here differential coding is used to reduce the bit rate needed for these parameters.

3. MODE DECISIONS

For each frame the mode decision is done in open-loop fashion: a set of parameters is extracted from the input signal prior to encoding, and used to classify the input frame. The mode decision method is described in this section.

3.1. Voice Activity Detection

A voice activity detector (VAD) is used to determine the presence or absence of speech in the current frame. We use a simple yet effective VAD which monitors variations in the energy level and the LPC spectrum. When the VAD indicates that the frame does not contain speech, the coder uses the Noise Mode.

3.2. Voicing Decision

When the VAD declares that speech is present, the open-loop pitch prediction gain, the zero-crossing rate, and a peakiness measure of the LPC residual are computed.

The open-loop pitch prediction gain and the zero-crossing rate are not sufficient for accurately classifying all speech frames. In particular, these measures are not able

to reliably detect the beginning or ending of a voiced utterance or detect unvoiced plosives which consist of short bursts of energy followed by a silent interval. The pulse codebook used in the Voiced/Transient Mode is well-suited for such events, because it can localize these bursts, whereas the stochastic codebook in the Unvoiced Mode tends to result in a spreading of the energy. Hence, we need to have a parameter which will help us classify such "transient" events into the Voiced/Transient Mode.

The peakiness measure is such a parameter. It is given by:

$$P = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N} r^2[n]}}{\frac{1}{N}\sum_{n=1}^{N} |r[n]|},$$

where $r[n]$ is the LPC residual and N is the frame size [3],[4]. If the signal contains a few pulses that are considerably larger in absolute value than the remaining samples, the peakiness measure is high; otherwise, it is low. Hence a large value of the peakiness measure occurs (1) for voiced speech, where the periodic pitch pulses can dominate the waveform, (2) at the start or end of a voiced segment, where a portion of the higher energy voiced signal is in the same frame as the lower energy unvoiced signal, and (3) for unvoiced plosives, which are characterized by a burst of energy, followed by a short silence.

If the pitch gain is low, the zero-crossing rate is high, and the value of the peakiness measure is low, the frame is classified as an unvoiced frame. Otherwise it is classified into either the Voiced/Transient or the Steady-State Voiced Mode, as discussed below.

3.3. Steady-State Voiced Mode Decision

A frame is classified into the Steady-State Voiced Mode if the following three conditions are all met:

1. The LPC envelope is changing slowly: the spectral envelope of the current frame is close to the previous frame's quantized spectrum in terms of weighted distance [5] in the line spectral frequency (LSF) domain.
2. The pitch is stationary: the pitch estimates for the current frame are near the coded pitch value of the last subframe of the previous frame.
3. The normalized open-loop pitch correlation is above a threshold dependent on the noise level estimate.

If any one of these conditions is not met, the frame is classified into the Voiced/Transient Mode.

4. CODING

The LPC parameters and the mode indicator are updated once per frame, while all remaining parameters are sent once per subframe. The 40 ms frame size presents some interesting challenges since the frame rate of the coder can at times be slower than the rate of change of the speech signal. These problems are addressed in several ways as described in the next few subsections.

4.1. Voiced Mode

4.1.1. LPC Vector Quantization

The LPC parameters are differentially quantized in the LSF domain with a first order autoregressive vector predictor and a 24-bit multi-stage vector quantizer (MSVQ) searched using an M-Best method with M=8 [6]. The predictor-quantizer pairs for each mode were designed separately using an optimal closed-loop procedure over the appropriate speech training sets. The predictors are diagonal matrix predictors. For the Voiced Mode, the predictor coefficients were about 0.7.

Because of the slow update rate, simply quantizing the LPC parameters once per frame is not sufficient for certain speech frames where the input spectrum changes rapidly. There is a need to represent more accurately the spectrum at intermediate points between the quantized vectors. In our coder this is achieved by a switched interpolation scheme: given the quantized LSF's of the current and previous frames, the LSF vector corresponding to their midpoint is interpolated using four different paths. The path for which the interpolated coefficients most closely match the original LSF vector is selected and transmitted using two additional bits.

4.1.2. Adaptive Codebook

The adaptive codebook lag is obtained with a two-step procedure. The first step consists of finding an open-loop estimate of the pitch over a large analysis window. The second step is a closed-loop refinement of the pitch value for each subframe [2].

For each 40 ms coding frame, two open-loop integer pitch estimates are calculated: the first one corresponding to the first 20 ms is called p_1, while the second one, corresponding to the second 20 ms is called p_2. Closed loop, fractional-pitch adaptive codebook search is performed for the first and third 10 ms subframes, in a small neighborhood around p_1 and p_2, respectively. These two values are encoded using 8 bits each. The pitch lags for the second and fourth subframes are searched in the vicinity of the closed loop lags for the first and third subframes, respectively, also using a closed-loop fractional-pitch adaptive codebook search procedure, and are differentially encoded using 4 bits each.

4.1.3. Fixed Excitation

The fixed codebook consists of sparse codevectors containing only four non-zero samples. These samples can have values of +1 or -1, and their locations are optimized using an M-Best search algorithm. The position and sign of each pulse are transmitted.

The coder uses a pitch sharpening mechanism which enhances the periodic nature of the speech signal. This method consists of introducing pitch periodicity into the pulse excitation and is turned on only when the pitch gain is high. An extra bit is used to signal to the decoder whether this option is turned on or off.

The adaptive and stochastic codebook gains are jointly vector-quantized in closed-loop fashion, by minimizing the perceptually weighted mean squared error between the original and coded speech waveforms.

4.2. Steady-State Voiced Mode

4.2.1. LPC Quantization

In the Steady-State Voiced mode, the LPC coefficients vary slowly. Hence the interframe vector predictor is much stronger (predictor coefficients above 0.9) than in the Voiced/Transient Mode, and we are able to use a

smaller, two-stage 12-bit MSVQ codebook. The codebook is once again searched with an M-Best procedure with M=8, and the switched interpolation described in Subsection 4.1.1 is also applied here.

4.2.2. Adaptive Codebook

The adaptive codebook search differs slightly from the one used in the Voiced/Transient Mode. In the Steady-State Voiced Mode, the open-loop pitch estimates p_1 and p_2 are both close to L_{last}, the closed-loop lag of the last subframe of the preceding frame. Hence, for pitch coding purposes, the open-loop estimate can be replaced by L_{last}. Thus, the pitch lags in the first and third subframes are coded differentially with respect to L_{last} using only 4 bits each. The coding of the pitch lags for the second and fourth subframes is done the same way as in the Voiced/ Transient Mode.

4.2.3. Fixed Excitation

In the Steady-State Voiced Mode, the signal is very strongly periodic, with a stable pitch period. For this reason, the adaptive codebook provides the dominant contribution to the quality of synthesized speech. Therefore, the fixed excitation contribution is reduced to 2 pulses per frame to lower the bit rate of this mode. Also, in this mode, the open loop pitch gain is always above the relevant threshold; Hence, the pitch sharpening is always turned on and the corresponding side information (1 bit) does not need to be transmitted.

4.3. Unvoiced Mode

4.3.1. LPC Quantization

In the Unvoiced Mode, the LPC parameters are predictively coded with a weak first-order predictor (with coefficients about 0.3-0.4) and a two-stage, 12-bit MSVQ codebook with an M-Best search and the switched parameter interpolation described above.

4.3.2. Gain-Matched Analysis-by-Synthesis

In this mode, there is no need for an adaptive codebook. Hence, only a fixed stochastic excitation codebook is used. A novel feature of our coder is the method used to code this excitation. This method helps overcome unwanted gain fluctuations, which often occur in CELP coders at low bit rates. This fluctuation is a result of the conventional analysis-by-synthesis coding method, where the stochastic codevector and its associated gain are chosen to minimize the weighted mean-squared error between the original and coded speech signals.

Let $\underline{s}$ be the perceptually weighted input speech signal minus the zero-input response of the weighted synthesis filter. Let H be the impulse response of the perceptually weighted LPC synthesis filter, $\underline{e}$ the excitation vector from the stochastic codebook, and g the associated gain term. Conventional CELP coders minimize the perceptually weighted mean-squared error (WMSE) between the original and coded speech, given by the expression

$$D = \|\underline{s} - gH\underline{e}\|^2$$

This expression needs to be minimized with respect to both g and $\underline{e}$. The optimal unquantized gain is obtained by setting to zero the derivative of D with respect to g:

$$g^{opt} = \frac{\langle \underline{s}, H\underline{e} \rangle}{\|H\underline{e}\|^2}$$

Substituting g^{opt} back into the expression for D, it can be shown that the optimal excitation is obtained by maximizing:

$$C = \frac{\langle \underline{s}, H\underline{e} \rangle^2}{\|H\underline{e}\|^2}$$

By observing the expressions for C and g^{opt}, it can be seen that, when the match is poor, as is the case in very low rate coders, a CELP coder tends to mute the gain value. This results in an annoying audible artifact in the coded speech, since perceptually, in most unvoiced signals, the gain of the excitation vector is more important than its exact time waveform. In our coder, we force the fixed excitation gain to be equal to the quantized version of the gain of the LPC residual of the input speech signal. The so-called target vector $\underline{s}$ used in CELP coding is divided by this gain factor to obtain a gain-normalized target vector $\underline{s}'$. The optimal excitation vector $\underline{e}$ is obtained by minimizing:

$$D' = \|\underline{s}' - H\underline{e}\|^2 \ ,$$

assuming that all candidate vectors have approximately unit norm. This is equivalent to maximizing:

$$C' = \|H\underline{e}\|^2 - 2\langle \underline{s}', H\underline{e} \rangle$$

This technique ensures that the synthetic speech has the correct gain value. At the same time, analysis-by-synthesis is still performed in order to help retain the character of the input signal. We call this method gain-matched analysis-by-synthesis. Listening tests have confirmed that this method considerably improves the perceived quality of unvoiced speech as well as background noise frames encoded with the Unvoiced Mode, as compared to the conventional CELP search method.

We use an overlapping codebook populated by center-clipped, zero-mean, unit-variance, Gaussian random numbers. This is a 7-bit codebook, with an additional sign bit, for a total of 8 bits.

4.4. Noise Mode

For background noise, as mentioned in Section 2, only the spectral envelope and the gain parameters are quantized. By definition, in this mode these parameters do not change rapidly. For this reason, we use strongly predictive quantization of these parameters at very low rates. However, low-rate quantizers may sometimes introduce rapid fluctuations in the quantizer output for such signals. This results in an output signal which can be characterized as “busy noise” or “swirling noise”. This problem can be rectified by a special quantizer designed to reduce the output signal fluctuations. If $X(n)$ is the original signal, and $Q(n-1)$ the quantized signal at the previous time instant, our implementation of reduced fluctuation quantization consists of obtaining a smoothed version of $X(n)$, called $X'(n)$, before quantization. This is done according to the equation:

$$X'(n) = (1-w)X(n) + wQ(n-1)$$

where w is a real number between 0.5 and 0.8.

4.5. Enhanced Adaptive Postfiltering

The decoder is equipped with an LPC-based adaptive postfilter, which improves the perceived speech quality by attenuating the quantization noise in the formant val-

leys and between pitch harmonics [7]. While a conventional postfilter works well for clean speech, it tends to create an unnatural speech quality in the presence of background noise. To overcome this problem, we use an enhanced postfilter where the amount of postfiltering is adapted according to an estimate of the signal to background noise ratio. If the ratio is high, postfiltering is performed as usual; if it is low, the strength of the postfiltering is gradually reduced by scaling down the filter bandwidth expansion coefficients.

5. BIT ALLOCATIONS

As previously mentioned, the coder has a frame size of 40 ms and 4 subframes per frame. Its bit allocations are summarized in Table 1.

Parameter	Noise	Unvoiced	Voiced/ Transient	Steady State
Mode Bits	2	2	2	2
LPC	12	12+2	24+2	12+2
Pitch Lags	0	0	8+4+8+4	4x4
Fixed Excitation	0	4x7	4x20	4x11
Excitation Signs	0	4x1	4x4	4x2
Gains	4	4x5	4x7	4x7
Pitch Sharpening	0	0	1	0
Total bits/frame	18	68	177	112
Rate (bits/s)	450	1700	4425	2800

Table 1: Bit Allocations

6. SUBJECTIVE LISTENING TESTS

We conducted subjective listening tests in our laboratory to evaluate the quality of the 3 kb/s coder (which we refer to as variable rate TI-CELP or VR-TI-CELP), in comparison with two reference coders: the 5.6 kb/s VSELP half-rate European cellular standard (GSM-HR), and the 13 kb/s RPE-LTP full-rate European cellular standard (GSM-FR). The test set contained 9 clean speech files (four females, four males and 1 child), and 11 files with various levels and types of acoustical background noise (6 females and 5 males). The background noise conditions ranged from mild to severe. All the files contained two or more sentences and about half of them came from a database of voice mail messages. The average bit rate over the 20 files was about 3 kb/s. For typical files the percentage of the time that each mode is used is roughly: 20% Noise, 12% Unvoiced, 45% Voiced/Transient, and 23% Steady State Voiced.

In the test, 15 naive listeners were presented with 40 speech sample pairs, where each pair consisted of the same file coded using two different coders. In each pair, one of the samples was always coded with VR-TI-CELP, while the other was coded with one of the two reference coders. The presentation order of the files in each pair as well as the sequence of the sentence pairs were both random. The listeners were not given any information about the coders in the test, and were asked to make a forced-choice preference in favor of one of the two coders for each pair. The results are summarized in Tables 2 and 3.

Test Condition	VR-TICELP (3 kb/s avg)	GSM-HR (5.6 kb/s)
Clean Speech	58%	42%
Noisy Speech	57%	43%
TOTAL	58%	42%

Table 2: AB test results (the numbers represent the percentage of the time that a coder was preferred)

Test Condition	VR-TICELP (3 kb/s avg)	GSM-FR (13 kb/s)
Clean Speech	71%	29%
Noisy Speech	52%	48%
TOTAL	60%	40%

Table 3: AB test results (the numbers represent the percentage of the time that a coder was preferred)

This listening test indicates that VR-TI-CELP is at least comparable to both GSM coders, but at a substantially lower average rate.

7. CONCLUSIONS

We have developed a variable-rate speech coder with low average bit rate and good quality. The coder uses several new techniques including the use of a peakiness measure for mode decisions and gain-matched analysis-by-synthesis coding. Over a database of voice-mail messages, the average rate of the coders was found to be about 3 kb/s and the speech quality was found to be at least equivalent to the GSM-HR and GSM-FR coders.

8. REFERENCES

[1] A. Das, E. Paksoy, A. Gersho "Multimode and Variable-Rate Coding of Speech", in *Speech Coding and Synthesis*, (W.B. Kleijn and K.K. Paliwal editors), pp.257-288,Elsevier 1995.

[2] W. LeBlanc, C. Liu, V. Viswanathan, "An Enhanced Full Rate Speech Coder for Digital Cellular Applications", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Volume 1, pp 569-572, 1996, Atlanta.

[3] D.L. Thomson and D.P. Prezas, "Selective Modeling of the LPC Residual During Unvoiced Frames: White Noise or Pulse Excitation", *IEEE International Conference on Acoustics, Speech and Signal Processing,* 1986, Tokyo, pp.3087-3090.

[4] A.V.McCree, T.P. Barnwell III, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding", *IEEE Transactions on Speech and Audio Processing,* Volume 3, Number 4, pp.242-250.

[5] K.K. Paliwal, B.S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Transactions on Speech and Audio Processing,* Volume 1, Number 1, pp.3-14.

[6] W.P. LeBlanc, B. Bhattacharya, S.A. Mahmoud, and V. Cuperman, "Efficient Search and Design Procedures for Robust Multi-Stage VQ of LPC Parameters for 4 kb/s Speech Coding", *IEEE Transactions on Speech and Audio Processing*, Vol. 1, Number 4, October 1993, pp. 373-385.

[7] J.-H. Chen, A. Gersho, "Adaptive Postfiltering for Quality Enhancement of Coded Speech", *IEEE Transactions on Speech and Audio Processing,*Volume 3, Number 1, pp.40-47.