A CELP VARIABLE RATE SPEECH CODEC WITH LOW AVERAGE RATE

Lei Zhang,¹ Tian Wang,² Vladimir Cuperman¹,²

¹School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada
²Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA

ABSTRACT

This paper presents a variable-rate CELP codec which achieves good communications speech quality at an average rate of about 3 kb/s. The codec operates as a source-controlled variable rate coder with rates of 4.9 kb/s for voiced and transition sounds, 3.0 kb/s for unvoiced sounds and 670 b/s for silent frames. New techniques used in the codec include prediction of the fixed codebook target vector and joint optimization of the adaptive and fixed codebook search. The prediction of the fixed codebook target vector is based on fixed codebook selections in previous subframes and a running estimate for the fundamental frequency. Informal subjective testing (MOS) indicates that the proposed codec, at an average rate of less than 3.2 kb/s, achieves better quality than fixed rate standard codecs with rates in the range 4 - 4.8 kb/s.

1. INTRODUCTION

In the past decade, Code Excited Linear Prediction (CELP) has become the dominant speech coding algorithm for bit rates between 4 kb/s and 16 kb/s. However, for rates around 4 kb/s and below, CELP loses its competitive edge to spectral domain coding. For many applications, an attractive approach for increasing system capacity while maintaining good quality is to allow the bit rate to vary according to the input speech characteristics. The corresponding codecs are characterized by their average bit rate and belong to the class of source-controlled variable rate codecs.

Source-controlled variable rate speech coders have been applied to digital cellular communications and to speech storage systems such as voice mail and voice response equipment. In both cases replacing the fixed-rate coders by variable-rate coders results in a significant increase in the system capacity while maintaining the desired quality of service.

Variable rate speech coders exploit two important characteristics of speech communications: the large amount of silence during conversation, and the large local changes in the minimal rate required to achieve a given speech reproduction quality. Studies have shown that the average speaker in a two-way conversation is talking about 36% of the time [1]. However, in this paper we consider one-way communications for which the speech activity factor increases from 36% to about 70%. This paper shows that even at this high speech activity factor, CELP-based codecs can achieve good quality at average rates as low as 3 kb/s.

Tutorial presentations of the variable rate speech coding can be found in [2-4]. Previous work on variable rate speech coding includes [5-10].

In this paper, we present a variable rate CELP codec (VR-CELP) with an average rate of about 3 kb/s. The codec uses a modular structure in which the general structure and coding algorithm is the same for all rates. All configurations are based on the system with the highest bit-rate. The lower bit-rates are obtained by varying the frame/subframe sizes, using different codebooks for quantization, and in some cases disabling codec components.

2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of the encoder. The codec operates as a source-controlled variable rate coder with rates of 4.9 kb/s for voiced and transition sounds, 3 kb/s for unvoiced sounds, and 670 b/s for silent frames. The appropriate coding rate is selected by analyzing each input speech frame using the frame classifier.

The codec uses standard techniques for computing and quantizing the LPC parameters represented as Line Spectral Pairs (LSPs). In order to reduce the rate for voiced and transition frames, the fundamental period (pitch) p is estimated and used to limit the range of the adaptive codebook indices used in the search.

The excitation signal for voiced/transition frames is formed as a summation of gain-scaled vectors from a fixed codebook, a predicted vector, and a three-tap adaptive codebook (ACB). The predicted vector is computed based on the fixed codebook selections in the previous subframes and the estimated pitch value p, and attempts to exploit the residual pitch-lag correlation in the fixed-codebook target vectors. This approach will be presented in detail in the next section.

The excitation signal for unvoiced frames is obtained from the fixed codebook, while disabling the adaptive codebook. All codebooks are disabled for silent frames, in which case a pseudo-random sequence known at both the encoder and the decoder is used.

Vectors in the adaptive and fixed codebooks and all gains are selected using an analysis-by-synthesis search based on a perceptually weighted MSE distortion criterion. Joint optimization of the ACB and the fixed codebook indices and closed-loop gain quantization is used in the search procedure.

In order to reproduce the excitation signal at the decoder, part or all of the following parameters are needed: class, quantized Line Spectral Pair (LSP) values, ACB center tap index, pitch value, fixed codebook indices, and quantized gains. Depending on the class information, the decoder duplicates the excitation signal, and passes it through the synthesis filter to obtain the reconstructed speech. The
such decision can not be made, the next parameter is used for
classification. If no parameter can classify the frame as
voiced or unvoiced, the frame is classified as a transition
frame.

The short-term predictor $1/A(z)$ is a tenth order LPC
all-pole filter. A perceptual weighting filter of the form
$H(z) = A(z)/A(z/\gamma)$ is derived from $A(z)$. Band-width
expansion and high-frequency compensation are used during
the LPC analysis. The LPC coefficients are computed once
per frame and converted to LSP values for quantization and
interpolation. The quantized LSPs are linearly interpolated
every subframe and converted back to LPCs to update the
synthesis filter. A tree-searched, multi-stage, vector quan-
tizer (MSVQ) [11] with four stages of 6 bits each for a total
of 24 bits is used for voiced and transition frames. After
each stage, the top three candidates which minimize the
weighted distortion criteria are retained. Unvoiced frames
use only the first two stages of the same MSVQ structure,
and silence frames use only the first stage.

We use an adaptive post-filter similar to that presented
in [12] which consists of a short-term pole-zero filter based
on the quantized short-term predictor coefficients followed
by an adaptive spectral tilt compensator. The pole-zero
filter is of the form $H(z) = A(z/\beta)/A(z/\alpha)$ where $\beta = 0.5$
and $\alpha = 0.8$. An automatic gain control is used to avoid
large gain excursions.

3. EXCITATION GENERATION AND
ENCODING

One of the problems typical of low-rate CELP codecs is
the residual pitch correlation which can be observed in the
fixed-codebook target vector. The pitch lag correlation can
not be modeled properly at the level of the fixed codebook
and this results in noisy reconstructed speech. This problem
can be observed at rates as high as 8 kb/s, and becomes a
predominant source of degradations at rates around 4 kb/s
or lower.

The reduced number of bits per subframe available at
rates around 4 kb/s leads to the use of limited-range ACB
search which is another potential source of increased pitch-
lag correlation for the fixed codebook target vector. Figure
2 illustrates a typical fixed-codebook target sequence (at
expanded scale) which shows strong correlation from one
pitch period to another, even after subtracting the ACB
contribution.

Figure 2. Fixed codebook target excitation

In order to alleviate this problem, a novel feature in-
troduced in this codec is a predicted fixed-codebook vec-
tor. This vector is obtained from the fixed-codebook
contributions of the previous subframe(s) as explained below.
The basic idea is to exploit the residual pitch-lag correla-
tion to improve the quality (without increasing the rate)
by using an additional contribution to the fixed-codebook
vector based on the fixed codebook entries from previous subframes.

For each subframe, the total fixed-codebook contribution to the excitation, $\xi_f$, can be written as

$$\xi_f = g_p \xi_p + g_f \xi_f,$$  \hspace{1cm} (1)

where $\xi_p$ and $g_p$ are respectively the predicted vector and its gain, and $\xi_f$ and $g_f$ are the fixed codebook vector and its gain. Note that a separate gain is introduced for the predicted vector, this gain is optimized in closed-loop, quantized, and transmitted to the receiver.

The fixed-codebook total excitation, $\xi_t$, is stored in a buffer $b$ of length $K$ using a procedure similar to that used in storing previous total excitation in the ACB buffer. The selection of the predicted vector $\xi_p$ for the next subframe can be viewed as sliding a window of length $N$, where $N$ is the subframe length, over the buffer $b$ to a position determined by the current pitch estimate. Figure 3 illustrates the process of selecting the predicted vector. The predicted vector can be expressed as

$$\xi_p(n) = \begin{cases} b(K - (p - n)) & n \leq p \quad \& \quad n \leq N \\ 0 & n > p \quad \& \quad n \leq N \end{cases}$$  \hspace{1cm} (2)

![Figure 3. Prediction vector computation](image)

The fixed codebook is based on a multipulse approach using only two pulses encoded with 9 bits for each subframe of 6 ms (48 samples). When the estimated pitch delay $p$ for the current subframe is less than the subframe size $N$, the selected fixed codebook vector $\xi_f$ is expressed as

$$c_f(n) = \begin{cases} \hat{\xi}_f(n) & n = 1,...p \\ \hat{\xi}_f(n-p) + \hat{\xi}_f(n) & n = p+1,...N \end{cases}$$  \hspace{1cm} (3)

where $\hat{\xi}_f$ is the codevector containing two pulses.

The adaptive codebook consists of total past excitation sequences. Only lags in a narrow window (4 samples) centered on the estimated pitch period value are considered in the ACB closed-loop search. Integer estimates for the pitch lag are obtained by an open-loop pitch estimator based on the SIFT method applied to the ideal excitation signal similar to that presented in [13]. A pitch tracker is used to reduce pitch doubling and pitch halving. The pitch value for each frame is encoded in 7 bits. ACB center tap index is coded in 2 bits (4 samples) for each subframe.

For voiced and transition subframes, the excitation parameters for the adaptive and fixed codebooks are determined in a closed-loop search which involves joint optimization of the adaptive codebook index, fixed codebook index, and all gain values. The joint-optimization minimizes the perceptually weighted MSE defined by

$$\epsilon = \| t - g_{a1} H \xi_{a1} - g_{a2} H \xi_{a2} - g_{a3} H \xi_{a3} - g_p H \xi_p - g_f H \xi_f \|^2$$  \hspace{1cm} (4)

where $t$ is the target vector (weighted speech vector after subtracting the weighted synthesis filter ZIR), $\xi_{a1}$, $g_{a1}$, $i = 1, 2, 3$, are the vectors and the gains for the 3-tap adaptive codebook, $H$ is the weighted synthesis filter impulse response matrix, and the other symbols were previously defined.

An exhaustive search is performed for minimization of (4) by computing the WMSE $\epsilon$ for all possible combinations of indices. This computational procedure is feasible due to the fact that there are only 4 possible ACB entries and the vector $\xi_p$ is fixed. For further complexity reduction, the vector $\xi_f$ can be split into two components according to its multipulse structure and the components can be searched sequentially. During the closed-loop search the gains in (4) are retrieved from the quantization tables resulting in the closed-loop quantization of the gains.

The unvoiced class uses only a fixed codebook containing sparse, ternary-valued, Gaussian white noise sequences. The contribution from the adaptive codebook and the predicted vector are set to zero. The fixed codebook index is coded with 8 bits for each subframe.

Both fixed and adaptive codebooks are omitted for silent frames. The excitation vector used to reproduce the background noise is obtained from a stochastic codebook using a pseudo-random index which can be identically generated at the encoder and the decoder.

4. CONFIGURATION SUMMARY AND CODEC PERFORMANCE

Table 1 gives the detailed bit allocation for the codec, including allocations for the short term predictor (STP), the adaptive codebook (ACB), and the fixed codebook (FCB) for each rate.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SL</th>
<th>UV</th>
<th>V/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame Size(samples)</td>
<td>144</td>
<td>144</td>
<td>288</td>
</tr>
<tr>
<td>RMS bits</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>STP bits</td>
<td>6</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>ACB bits</td>
<td>-</td>
<td>-</td>
<td>26</td>
</tr>
<tr>
<td>FCB bits</td>
<td>-</td>
<td>3x8</td>
<td>6x9</td>
</tr>
<tr>
<td>Gain bits</td>
<td>-</td>
<td>3x4</td>
<td>6x11</td>
</tr>
<tr>
<td>Class bits</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Bits/s</td>
<td>670</td>
<td>3000</td>
<td>4920</td>
</tr>
</tbody>
</table>

Table 1. Bit Allocations for Each Class

The effect of using fixed-codebook vector prediction is illustrated in Fig. 4 by comparing the reconstructed excitation with and without target vector prediction. This figure shows that the use of prediction results in a better match of the excitation than the conventional fixed code-book approach.

Figure 5 shows the frame by frame SNR of the reconstructed speech using prediction and joint search optimization, compared to the SNR obtained without using these two techniques. A significant improvement of the SNR is obtained for voiced frames and some transition frames.

Table 2 gives the results of a subjective quality evaluation of the CELP codec obtained by an informal mean opinion
Figure 4. Comparison of the excitation obtained with and without vector prediction. (a) fixed codebook target excitation; (b) \( g_{[EF]} \) fixed codebook reconstructed excitation without prediction; (c) \( g_{[EP]} + g_{[EF]} \) fixed codebook excitation with prediction.

score (MOS) test. The test was conducted with 20 participants (10 male, 10 female) listening to 8 sentences spoken by male and female speakers. Each file contained two sentences spoken by the same talker sampled at 8 kHz using 16-bit samples. The 4.1 kb/s IMBE standard and the DoD 4.8 kb/s CELP standard codecs were included as reference systems. Table 3 gives the classification mix generated by the variable rate system and the average bit rate.

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>MALE</th>
<th>FEMALE</th>
<th>BOTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>VR CELP</td>
<td>3.43</td>
<td>3.24</td>
<td>3.34</td>
</tr>
<tr>
<td>IMBE</td>
<td>2.97</td>
<td>3.16</td>
<td>3.07</td>
</tr>
<tr>
<td>DoD</td>
<td>3.04</td>
<td>3.03</td>
<td>3.03</td>
</tr>
</tbody>
</table>

Table 2. MOS Results

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>% V/T</th>
<th>% UV</th>
<th>% SIL</th>
<th>% BR (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>42.2</td>
<td>20.5</td>
<td>37.3</td>
<td>2928</td>
</tr>
<tr>
<td>Female</td>
<td>52.1</td>
<td>23.4</td>
<td>24.5</td>
<td>3431</td>
</tr>
<tr>
<td>Both</td>
<td>46.9</td>
<td>21.9</td>
<td>31.2</td>
<td>3176</td>
</tr>
</tbody>
</table>

Table 3. Class Statistics and Average Rate for MOS Files

The results in table 2 indicate that the variable-rate CELP (VR-CELP) achieved at an average rate lower than 3.2 kb/s better quality than fixed-rate codecs having significantly higher rate.

REFERENCES


