Voicing-Based Codebook in Low-Rate Wideband CELP Coding

Driss Guerchi\textsuperscript{1}, Tamer Rabie\textsuperscript{1}, and Abdelrhani Louzi\textsuperscript{2}

\textsuperscript{1}College of Information Technology, UAE University, Al Ain, UAE
\textsuperscript{2}Ericsson Canada, Montreal, Canada

guerchi@uaeu.ac.ae, tamer@uaeu.ac.ae, Abdelrani.louzi@ericsson.com

Abstract

In this paper we propose a new technique to quantize the spectral information in an Algebraic Code-Excited Linear Prediction (ACELP) wideband codec. The Voicing-Based Vector Quantization (VBVQ) presented in this paper, optimizes the search of the optimum vector while reducing the codebook entries by almost one third. In the VBVQ training phase, three codebooks are individually designed for voiced, unvoiced and transition speech. The proposed technique reduces the processing delay since it restricts the quantization of an input vector to only one of the three codebooks. For each speech frame, one codebook is selected based on the interframe correlation of the spectral information.

The VBVQ was successfully implemented in an ACELP wideband coder. The objective and subjective performance are superior to that of the combination of the split vector quantization and multistage vector quantization after using the same database for training and testing.

Index Terms: wideband speech coding, voicing-based vector quantization, low-complexity ISF quantization

1. Introduction

One of the critical attributes of the recent international speech coding standards is the computational complexity. The processing delay generated from the high complexity of a speech coder scheme might represent the major obstacle to its adoption by some real-time communication systems. New real-time applications, such as Voice over IP (VoIP), Internet telephony, video conferencing, video telephony, and wideband telephony, require high-quality speech coders with low processing delay. The ITU-T G.722.2 [1] wideband coder was recently borrowed to some of these applications since it meets most of the defined criteria. However, this standard, as well as other wideband codecs, still suffer from high computational complexity when implemented in remote conferences systems.

In order to minimize the analysis complexity, current speech coders ignore the interaction of the vocal tract shape with the vocal cords pitch. This suboptimal approximation is reflected in the disjoint operation of spectral quantization and pitch analysis. The quantization of the spectral parameters is done on a time frame basis without fully exploiting the interframe correlation with past frames. In order to model the spectral shaping filter with constant coefficients, the speech signal is considered as stationary during an analysis window. To achieve a fixed bit rate, the analysis window size is kept constant in most of the speech coding standards, regardless of the acoustical nature of the speech frame.

The optimal search of the best match to an input LPC (Linear Prediction Coding) vector is performed among all codewords of the codebook. While this method provides the best performance in terms of speech quality, its complexity is too high to be implemented in real-time applications. Several techniques have been proposed to reduce the computational complexity of the LPC quantization while sacrificing the speech quality performance [2]. The tree search and the multistage vector quantization are widely used to encode the spectral parameters. These approaches speed up the search procedure but also increase the memory requirements.

In this paper we introduce a new technique to reduce the search time for the optimal quantized LPC vector. The called Voicing-Based Vector Quantization (VBVQ) exploits the interframe correlation of the LPC parameters to limit the quantization process to a smaller codebook. The LP filter coefficients, as well as their different representations in the frequency domain, show some interframe redundancy. This redundancy, which is noticeable in voiced speech, could be used to predict the current LPC parameters from those of the past frames.

In the training phase of the voicing-based vector quantization, three disjoint ISF (Impittance Spectral Frequency; the LPC parameters from those of the past frames. As a consequence, the LPC quantization process in the VBVQ technique is preceded by the selection of the appropriate codebook. This selection is based on the interframe correlation of the current and previous LPC vectors. This approach not only provides high coded speech quality and reduced search complexity but also requires no extra bit resources.

2. Shortcomings of traditional LPC quantization

Most of the recent speech coder standards use Vector Quantization (VQ) to code the spectral information. While VQ techniques reduce bit rates, they increase the search computation load drastically. The performance of a VQ method is function of the size of the codebook. A codebook with more codewords certainly excels in coding efficiency the spectral parameters. This is achieved to the detriment of an increase in coding rate and computational complexity.

For example, in the G.729 narrowband codec standard [3], a combination of Multistage VQ (MVQ) and Split VQ (SVQ) is used to determine which 10-dimensional LSF (Line Spectral Frequency) vector corresponds most closely to the set of LSF input parameters. In the first stage of the search procedure, a codebook of 128 entries is searched; in the second stage two codebooks of 32 entries each are examined, for a total of 192 entries. In the G.722.2 wideband coding standard [4], the same VQ technique, with slight modifications, is employed to code 16 ISF coefficients. A total of 896
(256+256+128+128+64) entries are tested against the input ISF vector for all the codec modes, except for the 6.60 kbit/s coder which searches the closed codewords among 832 (256+256+128+128+64) entries. These numbers show the degree of the computational complexity of wideband coding, even when using suboptimal VQ techniques. A first step in easing the implementation of low-rate wideband speech coders in real-time communication applications, such as VoIP conferences, lies in minimizing the coder complexity [5]-[6].

An obvious remedy to the above problem consists of reducing the LPC quantization rate, which is related to the number of entries in the VQ codebooks. This solution will however deteriorate the coded speech quality. Another alternative is to confine the search of the closest codeword to a smaller number of codebook entries. This latter approach is widely used in the closed-loop pitch lag search, where an open-loop pitch analysis is performed first in order to limit the closed-loop pitch lag search to a few lags around the open-loop lags. Even though the objective is the same, conceptually the two approaches are different. The issue here is how to implement this idea without decreasing the codebook bit rate. We believe that in CELP coding, some aspects of the LPC quantization are still not fully utilized.

In high-pitched voices, a harmonic may coincide with a formant peak. In such situation, the LPC analysis filter will flatten the speech spectral envelope by removing not only the short-term relevant spectral details but also some energy of this harmonic. This phenomenon is illustrated in Figure 1. In this figure, the third harmonic of the speech spectrum coincides with the first formant. As can be seen in the LPC error spectrum, the energy of this harmonic is reduced after LPC analysis. We have come to the conclusion that some LPC parameters may bear pertinent information about the long-term periodicity in voiced speech. The current traditional LPC analysis and quantization ignore this LPC-pitch correlation since they are both performed on a time frame pitch-independent basis.

### 3. Voicing-based codebook

In this section we propose a new technique to reduce the computational complexity of the LPC quantization. This technique, called Voicing-Based Vector Quantization (VBVQ), consists of three disjoint codebooks; a voiced-speech codebook (VCB) quantizes the spectral information in voiced speech, an unvoiced-speech codebook (UCB) codes the unvoiced frames, and a third codebook (TCB) employed in transition speech. This technique looks similar to the split VQ. However, in the split VQ an input vector is divided into two or more subvectors. Each subvector is quantized using a separate codebook. In the VBVQ technique, one main virtual codebook is split into three codebooks, but each input vector is quantized using exclusively one codebook. The three codebooks are trained individually from voiced, unvoiced, and transition speech, respectively. The search of the closest match to an input LP filter vector is confined to only one codebook. This technique reduces the computational complexity without requiring any extra storage or bit resources. For each speech frame, LPC analysis is performed to extract 16 LP filter coefficients. The estimated LPC vector is compared (after conversion to an equivalent ISF vector) to the quantized LPC vector of the past frame using a squared error distortion measure. The selection of the optimal codebook for the current frame is based on the relative magnitude of this distortion. We expect that in voiced speech, consecutive LPC vectors are highly correlated. A small error distortion is a cue of quasi-stationary voiced speech.

We propose in this paper to exploit this correlation to estimate the current ISF vector from the past frame ISF coefficients. The motivation behind this idea is shown in Figure 2. The interframe variation of different formants is smooth in voiced segments, unlike in unvoiced speech where successive formants show very weak correlation. In the VBVQ algorithm, a 1st-order predictor, with coefficient 1, is applied to the input ISF vector. The residual ISF vector is quantized using exclusively either one of the VBVQ codebooks. The details of this algorithm will be given in the next section. Figure 3 illustrates the concept of the VBVQ.

![Figure 1](image1.png)

**Figure 1:** (a) Speech spectrum with the LPC spectrum superimposed; (b) spectrum of the LPC error signal.

![Figure 2](image2.png)

**Figure 2:** (a) Original wideband speech signal with voiced period marks; (b) Formant trajectory.

### 4. Voicing-based codebook design

In the VBVQ algorithm, three codebooks are trained from a large speech database. In the first phase of the training process,
we manipulate the speech database to build three sets of speech segments. The first set contains voiced speech, while the second and third are populated from transition and unvoiced speech, respectively. In the second phase, LPC analysis of order 16 is performed on 20ms speech frames from each set. The obtained LPC vectors are converted to ISF vectors. Three codebooks, VCB, TCB, and UCB using a combination of SVQ and MSVQ, are respectively designed for the three types of speech segments. The procedure is similar to the one used in the G.722.2 standard but with less bits for the ISF vectors quantization. To achieve a fixed bit rate, the three codebooks are allocated the same amount of bits (44 bits). It is worth noting that for the VCB codebook, a much smaller bit rate is sufficient to produce the same subjective performance as in the G.722.2 codec. The error between two consecutive LPC vectors is too small in voiced speech. Its variance allows significant bit rate reduction. At the end of the training process, each codebook will be characterized by one index i, from 0 to 2. Table 1 shows the bit allocation of the VBVQ method. r_n is the ISF error vector between the current ISF vector and the last frame quantized vector. In the first stage, this vector is split into two subvectors r_n,1 and r_n,2 of 9 and 7 coefficients, respectively. Then quantized to \( \tilde{r}_n,1 \) and \( \tilde{r}_n,2 \). In the second phase, the resulting quantization errors \( r_n - \tilde{r}_n,1 \) and \( \tilde{r}_n,2 \) are split into three subvectors and 2 subvectors, respectively. A total of 46 (44+2) bits is required. The two extra bits code the codebook index. A total of 640 entries are to be tested over the two stages of the quantization. This is a reduction of almost 30% compared to the combination MVQ-SVQ in the G.722.2 standard.

5. Selection of the optimal codebook

For every frame, the input ISF vector, \( p_n \), is compared to the last frame quantized ISF vector \( p_{n-1} \). A comparator checks the error distortion, \( r_n = p_n - p_{n-1} \), between the two vectors. If the energy of \( r_n \) is smaller than a certain threshold \( \epsilon_1 \), the VCB will be used for the search of the closest vector to the input vector \( p_n \). Otherwise UCB or TCB will be adopted according to the value of the square of \( r_n \) relative to another threshold \( \epsilon_2 \). The advantage of this method is that for steady-state speech frames, the percentage of hitting the optimum in the VCB codebook is greater than 95%. For non-stationary speech, such as unvoiced and transition between different phonemes, the TCB or the UCB will be used. Table 2 illustrates the algorithm for optimal codebook selection.

### Table 1: Bit Allocation of the VBVQ.

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_1 ) (7 bits)</td>
<td>(( r - r_1 )) (0 : 2) (6 bits)</td>
</tr>
<tr>
<td></td>
<td>(( r - r_1 )) (3 : 5) (7 bits)</td>
</tr>
<tr>
<td>( r_2 ) (7 bits)</td>
<td>(( r - r_2 )) (0 : 2) (5 bits)</td>
</tr>
<tr>
<td></td>
<td>(( r - r_2 )) (3 : 6) (5 bits)</td>
</tr>
</tbody>
</table>

### Table 2: Selection of the optimal codebook.

\[
\epsilon_n = \sum_{i=0}^{n} r_n^2(i)
\]

```plaintext
if \( \epsilon_n \leq \epsilon_1 \)
    optimal codebook = VCB
elseif \( \epsilon_1 < \epsilon_n \leq \epsilon_2 \)
    optimal codebook = TCB
elseif \( \epsilon_2 < \epsilon_n \)
    optimal codebook = UCB
end
```

6. Evaluation

We have conducted several simulations to compare the performance, in terms of objective and subjective measures, of the VBVQ to the G.722.2 SVQ-MVQ. The codebooks in both techniques have been trained using the same database. This is to avoid any effects of the selection of the database on the performance comparison. An informal listening comparative test has been performed as a subjective measure. As an objective measure, we select the Segmental Signal-to-Noise Ratio (SegSNR) at the output of the decoder. The systems to be evaluated are two versions of a wideband algebraic CELP coder. The two coders are similar except in the ISF quantization, where in coder 1 we use a combination of SVQ and MVQ. In the second ACELP coder, we implement the Voicing-based MVQ technique. The database for training phase consists of 150 min of English speech from 8 speakers; four women and four men. Each speaker read the same short utterance 10 times. We used the squared error ISP distortion for training and testing. However, the weighted distortion measure of Paliwal and Atal [7] is used to evaluate the ISF quantization in both versions of the ACELP coder. The evaluation simulations have been conducted on six different input sentences uttered by other speakers. In Table 3, we present the SegSNR for both ACELP versions. Table 4 shows the average spectral distortion between the input ISF vectors and their corresponding quantized ISF vectors.

7. Discussion and conclusions

As expected, the objective measure illustrates that the performance of the VBVQ approach is slightly superior to the combination of SVQ and MVQ whilst the complexity is reduced by almost one third in the VBVQ. The complexity of coding the other coder components, such as pitch lag, pitch gain and fixed codebook parameters, is unchanged since this paper addresses only the complexity generated from quantizing the spectral information. The informal listening tests reveal that the two techniques provide comparable overall subjective quality. For voiced speech, the VBVQ subjective performance is significantly higher. The VBVQ technique reduces the execution time of the ISF parameters quantization when operating at same bit rate as that of the G.722.2 SVQ-MVQ method. The overall complexity expense of the spectral parameters coding is reduced by almost 30%. However, this technique is not with
Table 3: Objective performance of the VBVQ technique.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>SegSNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVQ-MVQ</td>
</tr>
<tr>
<td>Female</td>
<td>10.65</td>
</tr>
<tr>
<td>Male</td>
<td>9.90</td>
</tr>
<tr>
<td>Average</td>
<td>10.275</td>
</tr>
</tbody>
</table>

Table 4: Spectral Distortion of the VBVQ technique.

<table>
<thead>
<tr>
<th>ISF Quantization</th>
<th>Avg SD (dB)</th>
<th>Outliers (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2-4 dB</td>
<td>&gt; 4 dB</td>
</tr>
<tr>
<td>SVQ-MVQ</td>
<td>1.31</td>
<td>2.41 0.06</td>
</tr>
<tr>
<td>VBVQ</td>
<td>1.19</td>
<td>2.44 0.08</td>
</tr>
</tbody>
</table>

no drawbacks, the efficiency of the VBVQ approach for highly noisy speech is affected. The correlation between two consecutive LPC vectors is not too high in noisy speech, even for voiced stationary segments. Unlike in clean speech, the number of wrong decisions in the selection of the optimal codebook increases in noisy voiced segments; an UCB might be selected to quantize the ISP vectors of a voiced frame. Since each of the three codebooks of the VBVQ technique is optimized specifically for only one type of speech (voiced, unvoiced or transition), a wrong decision in the selection of the optimal codebook will generate high quantization errors.

In future work, we plan on enhancing the performance of the VBVQ approach when speech is subject to background noise, and on increasing the robustness of the voicing decision.

8. References