AN 8 KBIT/S ACELP CODER WITH IMPROVED BACKGROUND NOISE PERFORMANCE

Roar Hagen and Erik Ekudden

Audio and Visual Technology Research
Ericsson Radio Systems AB
S-164 80 Stockholm, Sweden
{roar.hagen,erik.ekudden}@era.ericsson.se

ABSTRACT

This paper describes an 8 kbit/s ACELP speech coder with high performance for both speech and non-speech signals such as background noise. While the traditional waveform matching LPAS structure employed in many existing speech coders provides high quality for speech signals, it has significant performance limitations for e.g. background noise. The coder presented here employs a novel adaptive gain coding technique using energy matching in combination with a traditional waveform matching criterion providing high quality for both speech and background noise. The coder has a basic structure similar to that of the 7.4 kbit/s D-AMPS EFR coder, with a 10th order LPC, high resolution adaptive codebook and a 4-pulse algebraic codebook. The performance for speech signals is equivalent to or better than that of state-of-the-art 8 kbit/s coders, while for background noise conditions the performance is significantly improved.

1. INTRODUCTION

Speech coders at 7-13 kbit/s have found numerous applications in e.g. communication and voice storage systems. In fixed and mobile telephony, speech quality is of very high importance, and the service carries high user expectations. For mobile telephony, it is essential that a level of quality similar to that of fixed (wireline) situations can be provided. With the widespread use of mobile phones in all kinds of environments, such as in offices, in buses and in cars, and on the streets, the requirement for wireline speech quality extends also to these conditions. High quality also in the presence of various background noise types has become equally important to clean speech quality.

Current high-quality speech coders at bit-rates around 8 kbit/s are commonly using the Linear Prediction Analysis-by-Synthesis (LPAS) principle [1]. Two recently standardized coders using this principle are the 8 kbit/s G.729 Conjugate Structure - Algebraic Codebook Excited Linear Prediction (CS-ACELP) [2] and the 7.4 kbit/s D-AMPS Enhanced Full Rate ACELP [3], and represent state-of-the-art around 8 kbit/s. While several coders around 8 kbit/s provides wireline quality or near wireline quality for speech, significant performance losses are usually noted for speech in background noise and for background noise alone. Hence, the main goal of the current work has been to develop an 8 kbit/s coder that retain (or improve) the quality for clean speech, while significantly improving performance in background noise conditions.

Section 2 provides a background to LPAS coding and error criterion, Section 3 describes the new adaptive criterion, Section 4 gives a description of the complete 8 kbit/s coder, listening test results are given in Section 5, and finally Section 6 provides a summary.

2. BACKGROUND

Figure 1 shows the encoder structure of the CELP model using LPAS. One of the key elements of the LPAS principle is the minimization of the squared error criterion in a weighted speech domain. This criterion is given by

$$D_W = \left\| \mathbf{W}\cdot\mathbf{s} - \mathbf{W}\cdot\mathbf{H}\cdot\left(g_a\cdot\mathbf{c}_a + g_f\cdot\mathbf{c}_f\right)\right\|^2 \qquad (1)$$

where **W** and **H** are matrices performing the filtering operation of the weighting and synthesis filters. Using this criterion the uncoded speech vector **s** is compared with the coded speech signal generated from the adaptive and fixed codebook vectors $\mathbf{c}_a$ and $\mathbf{c}_f$ and their associated gain factors g_a and g_f. Especially for voiced speech, this criterion and the efficient adaptive codebook provide good speech quality in the bit-rate range of 7-13 kbit/s.

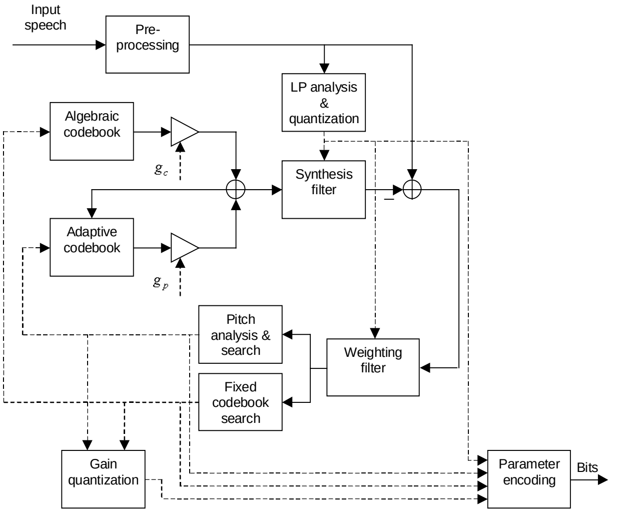

Figure 1. Principle of CELP encoder with LPAS.

For noise-like segments, such as unvoiced speech and background noise, the criterion (1) is less efficient and noticeable artifacts are introduced. The adaptive codebook is less efficient due to the lack of long-term periodicity. Thus, the waveform matching abilities of the coder are not good enough leading to a coded signal with a too low level. The optimal uncoded gain value, g, for a given codebook vector is given by:

$$g = \frac{\mathbf{x}^t \cdot \mathbf{c}}{\mathbf{c}^t \cdot \mathbf{c}} \qquad (2)$$

where $\mathbf{c}$ is the codebook vector used to match the target vector $\mathbf{x}$. From eq. (2) it is seen that the gain is given by the waveform matching through the cross-correlation in the nominator. Thus, poor matching leads too a low gain value.

For stationary background noise, these shortcomings also manifest themselves in an artifact known as swirling [4]. An unnatural time-varying sound is perceived which might be partly due to varying waveform matching abilities (it is also believed to be due to fluctuations in parameter estimates of the linear prediction filter). This is supported by the observation that the coding gain of the adaptive codebook fluctuates strongly in stationary background noise. Thus, according to eq. (2), some segments have high gain values whereas other segments have low values leading to a time-varying sound. In Figure 2, the coding gain (as computed by eq. (5)) of the adaptive codebook is depicted for a speech segment with street noise in the background.

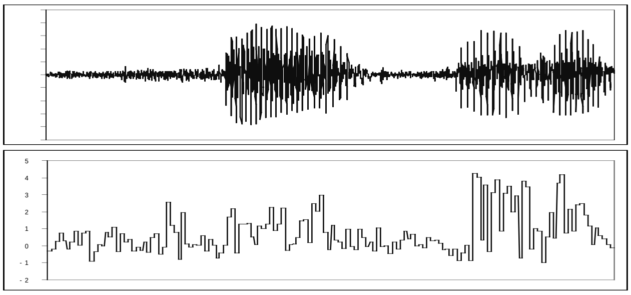

Figure 2. Top: Speech signal. Bottom: Coding gain for the adaptive codebook.

It is known in the literature that noise-like segments have other perceptual requirements than voiced segments in terms of the important features to reproduce. Multi-mode and variable-rate coders have exploited this by changing the coding strategy depending on e.g. a voicing classification procedure. A recent example is found in [5] where, for the unvoiced mode, the adaptive codebook is not used. Furthermore, the fact that the gain is more important than the exact waveform match is exploited by setting the gain of the fixed codebook to a value so that the energy level of the LP residual is matched instead of using eq. (2).

The coder presented here is a single-mode coder and uses an adaptive codebook irrespective of the character of the speech segment. An experiment was performed for the purpose of evaluating whether (i) the criterion (1) is causing some of the artifacts in noise-like segments; and (ii) energy matching decreases the artifacts. In the experiment, after encoding a subframe, the fixed codebook gain was re-computed so that the excitation signal has the same energy as the LP residual. Both the adaptive and fixed codebook gains were unquantized. Informal listening tests demonstrated that the annoying artifacts in background noise were eliminated. However, for voiced sounds new artifacts were introduced giving a noisy character to the synthesized speech. In conclusion, this experiment verified the weakness of the waveform matching criterion (1) in noise-like segments as well as the fact that energy matching is important.

3. A NEW ADAPTIVE CRITERION

The knowledge gained by the experiment described in the previous section also suggested a way to exploit the importance of energy matching in noise-like segments without retreating to a multi-mode approach with its drawbacks. We formulated a new adaptive criterion to be used for the encoding process. This criterion is given by:

$$D = (1-\alpha)\cdot D_W + \alpha \cdot \left(\|\mathbf{s}_w\| - \|\tilde{\mathbf{s}}_w\|\right)^2 , \qquad (3)$$

where $\mathbf{s}_w = \mathbf{W}\cdot\mathbf{s}$ is the weighted speech signal and $\tilde{\mathbf{s}}_w$ is the weighted synthesized speech signal given by:

$$\tilde{\mathbf{s}}_w = \mathbf{W}\cdot\mathbf{H}\cdot\left(g_a\cdot\mathbf{c}_a + g_f\cdot\mathbf{c}_f\right). \qquad (4)$$

Equation (3) gives a mixture of waveform matching and energy matching through the adaptive balance factor α . This criterion has several advantages as compared to a conventional solution:

- Waveform matching and energy matching is softly combined in order not to rely on either one or the other.
- Through the adaptive nature of the criterion, the balance can be smoothly adjusted over time to avoid drastically changing coding strategy.
- Some waveform matching can always be maintained.

While it is possible to use the new criterion for the entire encoding process in a LPAS coder, we have used it only for gain quantization. Therefore, criterion (1) is used to find the adaptive and fixed codebook vectors. This is advantageous for implementation purposes, makes adaptation easier, and has achieved the improvements demonstrated by the initial experiment.

3.1 Adaptation

A key to the new criterion is the adaptation through the balance factor α . For good performance, it is crucial that a suitable value of α is used at each subframe. Since the criterion (1) is tied to the effectiveness of the adaptive codebook, we base the adaptation on the coding gain for the adaptive codebook:

$$v = 10\cdot\log\frac{\|\mathbf{r}\|^2}{\|\mathbf{r}-g_a\cdot\mathbf{c}_a\|^2} , \qquad (5)$$

where $\mathbf{r}$ is the LP residual. The optimal unquantized value of the gain (cf. eq. (2)) is used. Note that a measure in the weighted speech domain can also be used. Due to the fluctuation of v as demonstrated in Figure 2, a median filtering of the coding gain is performed, $v_m =$ median($v, v_{-1}, \ldots, v_{-p}$) where the subscript denotes the coding gain of previous subframes. The adaptation factor is now given as a function of v_m :

$$\alpha = f(v_m) = \begin{cases} c & v_m < a \\ c + \dfrac{d-c}{b-a}\cdot(v_m - a) & a \le v_m \le b \\ d & v_m > b \end{cases}. \qquad (6)$$

The parameters are set so that a maximum value c is used for a coding gain below a, a minimum value d is used for a coding gain above b, and a linear transition region between.

Due to the importance of waveform matching at onsets, a simple onset detector was used to force the α value to its minimum at onsets.

The median filtering introduces smoothness over time in the α values. However, it was found beneficial to introduce an additional averaging when the α value moves out of the upper saturation region.

3.2 Vector Quantization of Gains

For vector quantization (VQ) of the gains, the procedure outlined above is straightforward. The criterion (1) is used for all encoding steps except for the search of the gain VQ codebook, where the new criterion (3) is used.

3.3 Scalar Quantization of Gains

For scalar gain quantization, the conventional approach is to use a squared error criterion in the gain domain where the possible quantized values are compared to the optimal unquantized value as given by eq. (2). Since the fixed codebook gain is the important parameter for noise-like segments, we employed the conventional criterion for the adaptive codebook gain whereas the new adaptive criterion is used for the fixed codebook gain. The criterion (3) was reformulated for this purpose according to

$$D_{SQ} = (1-\alpha)\cdot\|\mathbf{c}\|^2\cdot(g-\hat{g})^2 + \alpha\cdot\left(\|\mathbf{r}\| - \hat{g}\cdot\|\mathbf{c}\|\right)^2 \qquad (7)$$

where $\hat{g}$ denotes the quantized gain. The scalar gain codebook is searched using eq. (7).

4. CODER DESCRIPTION

The coder is based on the D-AMPS EFR coder [3] using 20 ms frames, 5ms subframes and 5 ms lookahead. The bit-allocation of the 8 kbit/s coder is shown in Table I. Scalar gain quantization is used as better performance was obtained with the new criterion than when using a VQ. Furthermore, bits have been added to the LP filter and adaptive codebook to increase clean speech performance.

Table I. Bit allocation for the 8 kbit/s ACELP.

Parameter	Subframe 1&3	Subframe 2&4	Total
LP coeff.			27
Pitch parity			1
Adapt CB index	8	6	28
Adapt CB gain	4	4	16
Alg CB index	13	13	52
Alg CB sign	4	4	16
Alg CB gain	5	5	20
Total			160

4.1 LP analysis and quantization

A 10th order LP analysis is performed using the Levinson-Durbin algorithm. The autocorrelation function is computed from the windowed speech signal. The window is a hybrid Hamming-Cosine window of length 240 samples. Bandwidth expansion of 60 Hz as well as white-noise correction at –40 dB is applied to the autocorrelation function.

The resulting LP coefficients are converted to Line Spectrum Frequencies (LSFs) prior to quantization. A 1st order MA prediction is used to predict the LSFs of the current frame. The prediction residual is quantized using split VQ with subvectors of dimension 3, 3, and 4 with 9-bit codebooks for each subvector.

The perceptual weighting filter is computed from the unquantized LP coefficients in the same way as for the D-AMPS EFR coder [3].

4.2 Adaptive codebook

Twice per frame, an open-loop pitch analysis is performed in order to reduce the search complexity in the adaptive codebook. An open-loop pitch delay, T_o, is estimated from the weighted speech signal (the speech signal filtered by the perceptual weighting filter).

In the 1st and 3rd subframe, the adaptive codebook uses an 8-bit absolute coded pitch delay with a fractional resolution of 1/3 in the range [19 1/3, 84 2/3] and integer values from 85 to 143. The open-loop estimate is used to restrict the search.

In the 2nd and 4th subframe, the adaptive codebook uses a 6-bit delta-coded pitch delay. The delay is coded relative to the pitch delay T_1 of the 1st and 3rd subframe rounded to integer resolution. Fractional pitch with resolution 1/3 is used in the entire range [T_1-10 2/3, T_1+9 2/3].

4.3 Algebraic codebook

The fixed codebook employs the algebraic structure with 4 signed pulses in 4 non-overlapping tracks. Table II shows the track table for the algebraic codebook. Each pulse has one-bit sign giving a total of 4 sign bits. Pulse 1, 2, and 3 can take on one of 8 positions whereas pulse 4 is located at one of 16 positions. This gives 3, 3, 3, and 4 bits for position coding, a total of 13 position bits. The structure and search of the algebraic codebook is the same as in the D-AMPS EFR coder [3].

Table II. Track table for the algebraic codebook.

Pulse	Positions
1	0, 5, 10, 15, 20, 25, 30, 35
2	1, 6, 11, 16, 21, 26, 31, 36
3	2, 7, 12, 17, 22, 27, 32, 37
4	3, 8, 13, 18, 23, 28, 33, 38, 4, 9, 14, 19, 24, 29, 34, 39

4.4 Gain quantization

The adaptive codebook gain is quantized with a 4-bit non-uniform scalar quantizer according to the conventional criterion. The fixed codebook gain is quantized using the new criterion of eq. (7) with the adaptation of eq. (4-5). A 5-bit codebook is used.

4.5 Post-processing

Anti-sparseness processing

Due to the sparse algebraic codebook with 4 pulses per 40 samples subframe, a novel anti-sparseness processing [6] of the fixed codebook signal is performed. The annoying artifacts caused by the sparseness are removed by this procedure. These artifacts are most prominent for noise-like signal segments such as background noise. For such sounds, stronger anti-sparseness modifications are needed than for periodic speech segments where the adaptive codebook provides most of the excitation. Therefore, the impulse response characteristics are adapted to the local character of the speech in a similar way as the adaptation of the criterion described earlier. Since post-processing is done in the decoder, the adaptation is based on the adaptive codebook gain factor instead of the coding gain of the adaptive codebook. One of three impulse responses performing 1) strong, 2) medium, and 3) no modification is used. The impulse responses are adaptively selected according to the following procedure:

1. Select impulse response 1 if g_a<0.6, select impulse response 2 if g_a is in the range 0.6 to 0.9, select impulse response 3 if g_a>0.9.
2. Compute an onset indicator that is set if the current fixed codebook gain is more than twice the previous fixed codebook gain.
3. If the impulse response is not 1 and onset is not indicated, compute median filtered value of current g_a and the previous 4 values. If the result is less than 0.6, select impulse response 1.
4. If onset is indicated and the impulse response is not 3, increment the impulse response selected by 1.

This adaptation algorithm performs well and manages to use the impulse response with strong modification for pure background noise while working well for the speech segments. Since the adaptation is based on the quantized gain values, no extra information is needed to select the correct impulse response.

Post-filtering

Adaptive post-filtering including pitch and formant postfiltering is used. The postfiltering is identical to that of the D-AMPS EFR coder [3].

5. RESULTS

An informal A-B listening test was conducted involving 8 listeners. The subjects had to make a forced choice between pairs of samples presented over regular telephone handsets. The new 8 kbit/s coder was compared to G.729 for clean speech and speech with 15 dB car and street noise. The test material included 8 clean speech sentences and 5 sentences each with street and car noise. The sentence pairs were presented randomly and in both A-B and B-A order. The results are shown in Table III.

Table III. Subjective listening results (preference scores).

Condition	Clean	Street	Car	Total
G.729	44%	30%	22%	34%
8 kbit/s	56%	70%	78%	66%

As seen in the table, the new criterion (with the anti-sparseness processing) clearly improves performance for background noise conditions. The clean speech performance is equivalent to that of G.729 (although a slight, not statistically significant, preference for the new coder is seen)

6. SUMMARY

Lower bit-rate LPAS coders usually suffer from quality problems under background noise conditions. One of the reasons, within the constraints of the LP-model, is the strict waveform matching criterion employed. The novel adaptive criterion combines the traditional waveform matching criterion which provides high quality for clean speech with a new energy matching criterion which is more suitable for coding of noise-like signals, such as background noise. The new criterion together with an adaptive anti-sparseness post-processing technique has been implemented in an 8 kbit/s ACELP coder. The quality of the proposed 8 kbit/s coder has been demonstrated to be equivalent to that of a state-of-the-art 8 kbit/s coder for clean speech and significantly improved for background noise conditions. The new adaptive criterion can also be applied in existing coders to enhance performance for background noise.

REFERENCES

[1] W.B. Kleijn and K.K. Paliwal, *Speech coding and synthesis*. Amsterdam, Holland: Elsevier, 1995.

[2] R. Salami *et al*., "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder", *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 116-130, 1998.

[3] T. Honkanen, J. Vainio, K. Järvinen, P. Haavisto, "Enhanced full rate speech codec for IS-136 digital cellular system", in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Munich, Germany, pp. 731-734, 1997.

[4] T. Wigren *et al*., "Improvements of background sound coding in linear predictive speech coders", in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Detroit, MI, pp. 25-28, 1995.

[5] E. Paksoy, A. McCree, and V. Viswanathan, "A variable-rate multimodal speech coder with gain-matched analysis-by-synthesis", in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Munich, Germany, pp. 751-754, 1997.

[6] R. Hagen, E. Ekudden, B. Johansson, and W.B. Kleijn., "Removal of sparse-excitation artifacts in CELP", in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Seattle, WA, pp. I-145-148, 1998.