CLOSED LOOP DYNAMIC BIT ALLOCATION FOR EXCITATION PARAMETERS IN ANALYSIS-BY-SYNTHESIS SPEECH CODEC

James P. Ashley and *Udar Mittal*
ashley,mittal@labs.mot.com
Speech Processing Research Lab, Motorola Labs,
Schaumburg, IL-60196, USA

ABSTRACT

A method for dynamically allocating bits for the adaptive and fixed codebook of an analysis-by-synthesis speech codec is proposed. The bit allocation is based on the closed loop weighted mean squared error. The different bit allocations identify various codebook configurations used by the adaptive and fixed codebooks of the codec. Unlike open loop approaches where the decision on a codebook configuration is made once per frame, in the closed loop approach the decision on which codebook configuration should be used is made in each subframe. A variable length code is used for coding the codebook configuration. The factorial packing codebook is used as the fixed codebook. The technique is a part of 8.5 kbps mode of EVRC-WB speech coding standard.

Index Terms— CELP, adaptive codebook, fixed codebook, factorial packing.

1. INTRODUCTION

The synthetic excitation vector in Code Excited Linear Predictive (CELP) speech codec is the gain scaled sum of an adaptive codebook (ACB) excitation and a fixed codebook (FCB) excitation. These codebooks are searched to find the synthetic excitation which minimizes the weighted mean squared distortion between input speech and synthesized speech. The synthesized speech is the generated by passing the synthetic excitation through a linear prediction coefficient (LPC) filter. The ACB excitation is obtained from the past excitations using open loop pitch delay and searching for an optimum delay adjustment factor [1]. The FCB excitation is obtained from searching a stochastic codebook.

The open loop analysis procedure generates parameters such as LPC, and open loop delay once per speech frame. The closed loop method is used to obtain the ACB excitation and the FCB excitation in each subframe. The contribution of ACB excitation and the FCB excitation to the total synthetic excitation may vary depending on the characteristic of the speech frame.

Speech frames can be classified as voiced, unvoiced or transient. In voiced speech frames, the energy of the ACB excitation dominates the synthetic excitation where as in the transient or unvoiced speech frames the energy of fixed codebook dominates the synthetic excitation. Variable-rate speech codecs such as [1] use this property of speech to code voiced frames at half the bit-rate by allocating a lesser number of bits for the fixed codebook.

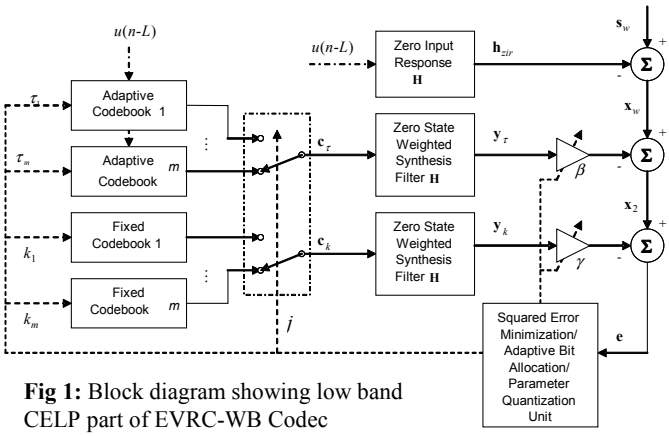

Fig 1: Block diagram showing low band CELP part of EVRC-WB Codec

Speech codecs proposed in [2,3] classify speech frames on the basis that a pitch delay is predictable for stationary voiced and hence use less bits for pitch delay (adaptive codebook) in such frames and use more ACB bits where pitch delay is not predictable.

The speech codecs described above allocate bits for the ACB and the FCB based on the open loop parameters only. Thus, the bits for both the codebooks in all the subframes are allocated based on initial classification of a speech frame. These codecs do not take into account that the transition from voiced to unvoiced or from transient voiced to stationary voiced can take place inside a frame. Hence, for some of the subframes in a given frame, the ACB may require more bits while other may require less ACB bits. Classifying each of the subframes based only on open loop analysis may adversely affect speech quality.

In this paper, we propose a method for dynamic bit allocation between the ACB and the FCB wherein the bit allocation is performed every subframe using the closed loop weighted mean squared error.

Notations: Let N be the frame length, $\mathbf{s_w}=\{s_1,s_2,\ldots, s_N\}$ be the vector representing the weighted speech signal for the current frame. Let L be the subframe length and $\mathbf{x_w}=\{w_1,w_2,\ldots, w_L\}^{\mathbf{T}}$ be the L dimension vector representing the weighted speech vector from which the zero input response of the weighted synthesis filter ($\mathbf{H}$) has been subtracted. Let $\mathbf{c_\tau}$ and $\mathbf{c_k}$ be ACB and FCB excitation vectors, and β and γ be their respective gains. The overall synthetic excitation, $\mathbf{c}$, is given by $\mathbf{c} = \beta\cdot\mathbf{c_\tau} + \gamma\cdot\mathbf{c_k}$. The other notations are shown in Fig. 1.

2. BIT ALLOCATION FOR EXCITATION PARAMETER

The dynamic bit allocation technique is integrated in the subframe loop of an EVRC-WB [1] speech coding standard which is a split band wideband codec. The split band wideband speech codec uses a CELP codec in the low band [0-4000 Hz]. The block diagram of the low band part of the codec is shown in Fig. 1. Each frame consists of 160 samples (20 ms), and is divided into three subframes of lengths 53, 53, and 54 samples. The synthetic excitation is obtained by minimizing the weighted mean square distortion between the RCELP [1,5] modified speech and the synthetic speech.

The ACB is populated by past excitations and is referenced by a delay contour[1]. The delay contour (τ) for a subframe is obtained from the previous frame's open loop pitch delay (τ_p), the current frames's open loop pitch delay (τ_c), and a delay adjust parameter δ. The optimum delay adjust parameter δ^* is typically obtained as:

$$\delta^* = \arg\max_{\delta\in\Delta}\left\{\frac{\left(\mathbf{x}_w^T\mathbf{H}\mathbf{c}_\tau\right)^2}{\mathbf{c}_\tau^T\mathbf{H}^T\mathbf{H}\mathbf{c}_\tau}\right\} \quad (1)$$

where set Δ contains possible values of the delay adjust parameter.

In non-dynamic bit allocation methods, the delay adjust parameter is searched over a single set Δ, which is pre-decided. On the other hand, in the proposed dynamic bit allocation method, the delay adjusts parameter is searched over more than one Δ, and the decision on which set to be used for obtaining δ^* is based on the weighted mean squared error. Details on this decision making procedure will be provided in Section 2.1. Let $\Lambda=\{\Delta_1,\Delta_2,\ldots,\Delta_m\}$ be the m sets which will be used for searching the delay adjust parameter. These m sets are also referred as the m sets of ACBs in Fig. 1. The number of bits (n_τ) needed to represent τ in all the three subframes is the sum of bits needed to represent delay adjust parameters, which is dependent on the cardinality of the set Δ. Thus, n_τ can be increased by selecting Δ_2 instead of Δ_1 for searching the delay adjust parameter.

The fixed codebook is a factorial packing codebook (FPC) [4], i.e., $\mathbf{c_k}$ consists of unit magnitude pulses whose sum is a constant, N_p. The number of bits n_k needed to accurately represent $\mathbf{c_k}$ is dependent on number of pulses [4], N_p, and is given by:

$$n_k = \left\lceil \log_2\left(\prod_{j=1}^{3}\sum_{i=1}^{\min(N_p(j),L_j)} 2^i\cdot\frac{n!}{i!(n-i)!}\cdot\frac{(Np(j)-1)!}{(i-1)!(Np(j)-i)!}\right)\right\rceil, \quad (2)$$

where j=1, 2, 3 is the subframe number, $N_p(j)$ is the number of pulses in the j-th subframe and L_1=53, L_2=53, *and* L_3=54 are subframe lengths. Note that for fixed subframe lengths, the FCB size decreases with decrease in N_p. By selecting δ^* from a larger ACB (set Δ with higher cardinality) and using a lower value of N_p for the FCB, the FCB bits can be traded for ACB bits. Let $\Omega=\{N_p^1,N_p^2,\ldots,N_p^m\}$. The values of N_p are decreasing in the above set. The set Λ and Ω are complementary in a sense that if delay adjust is selected from a set Δ_j then the FCB excitation is searched from a FCB having N_p^j pulses. A pair (Δ_j, N_p^j) will be referred as the j-th codebook configuration and the codebook structure consists of m codebook configurations.

2.1. Selection of the Codebook Configuration

If the delay adjust parameter is obtained by searching over a single adaptive codebook or a single set Δ, the expression in Eq. 1 is maximized during the adaptive codebook search. In the current method, we also need to decide from which adaptive codebook should δ belong. In the ideal sense, ($\mathbf{c_\tau}$, $\mathbf{c_k}$) should be searched from all codebook configurations (Δ_j, N_p^j) and then the synthetic excitation should be selected from the configurations which result in the minimum distortion. However, this can be prohibitively complex due to multiple FCB search iterations.

In a reduced complexity approach, we first define

$$\varepsilon_j^* = \min_{\delta\in\Delta_j}\left\{\left(\mathbf{x}_w^T-\beta\mathbf{H}\mathbf{c}_\tau\right)^2\right\}, \quad (3)$$

where

$$\beta = \max\left(0,\min\left(1.2,\frac{\left(\mathbf{x}_w^T\mathbf{H}\mathbf{c}_\tau\right)}{\mathbf{c}_\tau^T\mathbf{H}^T\mathbf{H}\mathbf{c}_\tau}\right)\right), \quad (4)$$

as the minimum distortion for the j-th ACB codebook vector, and then select ACB excitation from a codebook configuration resulting in minimum value of ε_j^*, i.e.,

$$\Delta = \arg\min_{\Delta\in\Lambda}\left\{\varepsilon_1^*,\varepsilon_2^*,\ldots,\varepsilon_m^*\right\}. \quad (5)$$

The problem in this approach is that if the ACB excitation is selected from a set Δ with higher cardinality, i.e, more ACB bits, then $\mathbf{c_k}$ will be searched from a smaller FCB. The gain from using more bits for ACB may not be sufficient to compensate for the loss from using a smaller

FCB. To account for this loss, bias factors are applied to ε_j^* in (5), i.e.,

$$\Delta = \arg\min_{\Delta \in \Lambda} \{\varepsilon_1^*, b_2\varepsilon_2^*, ..., b_m\varepsilon_m^*\}, \quad (6)$$

where $b_j > b_{j-1} > 1$ (Note that b_1=1). Even though use of b_j enables selection of larger size ACB when the optimum weighted distortion ε_j^* of that ACB is significantly lower, it does not account that ε_j^* may be large for all ACBs, i.e., the contribution from the ACBs is insignificant, and hence the larger size ACB will not be able to compensate for the loss from using a smaller FCB. This can be accounted for by computing the long-term prediction gain, defined as:

$$\lambda_j = \frac{\|\mathbf{x}_w\|^2}{\left\|\mathbf{x}_w - \beta \mathbf{H}\mathbf{c}_{\tau_j^*}\right\|^2}, \quad (7)$$

whenever any other codebook configuration (Δ_j, N_p^j) other than (Δ_1, N_p^1) is selected in (6). If the long term prediction gain is greater than a threshold λ_{th}, then the configuration (Δ_j, N_p^j) is selected.

3. ENCODING OF EXCITATION CODEBOOK

We now consider the codebook configuration which is part of the EVRC-WB standard. The configuration consists of two sets of codebooks. For the first set $\Delta_1 = \{0\}$ and $N_p^1 = 6$, and for the second set $\Delta_2 = \{-2\delta_{adj}, -\delta_{adj}, \delta_{adj}, 2\delta_{adj}\}$ and $N_p^2 = 5$, where δ_{adj} is a function of current and previous frames open loop delays. Note that in the first configuration there is no need for closed loop delay adjust search. The codebook configuration selection is performed in every subframe, hence each subframe may use a different codebook configuration. To code the excitation parameters using a minimum of bits, the codebook configuration, ACB excitation, and FCB excitation for the complete frame are coded together. A Huffman code is used for coding the codebook configurations for the three subframes. The encoding bit allocation is shown in Table 1. FPC is used for coding the fixed codebook. The bits used for FPC coding in Table 1 are calculated using equation (2). Two ACB (delay adjust) bits are used in the subframes using Configuration 2 and zero ACB bits are used for Configuration 1.

4. RESULTS

The total weighted signal-to-noise ratio (WSNR) and average WSNR defined as

$$\text{WSNR}_{\text{Tot}} = 10 \cdot \log\left\{\frac{\sum \|\mathbf{s}_w\|^2}{\sum \|\mathbf{e}\|^2}\right\}, \quad (8)$$

and

$$\text{WSNR}_{\text{Avg}} = \frac{1}{M}\sum 10 \cdot \log\left\{\frac{\|\mathbf{s}_w\|^2}{\|\mathbf{e}\|^2}\right\}, \quad (9)$$

respectively, compare the performance of the proposed dynamic bit allocation method to the fixed bit allocation methods. The summations in Equation (8) and (9) are over all speech subframes using CELP, and *M* is the total number of subframes.

The proposed method is compared to the following fixed ACB and FCB codebook structures:

1. Structure-1: $\Delta = \{0\}$ and $N_p = 6$, i.e., 0 bits for ACB and 93 bits for FCB.
2. Structure-2: $\Delta = \{0, -2\delta_{adj}, -\delta_{adj}, \delta_{adj}, 2\delta_{adj}\}$ and $N_p = 5$, i.e., 7 bits for ACB and 81 bits for FCB.
3. Structure-3: $\Delta = \{0, -2\delta_{adj}, -\delta_{adj}, \delta_{adj}, 2\delta_{adj}\}$ and $N_p = 6$, i.e., 7 bits for ACB and 93 bits for FCB.

In each of the above codebook structures as well as the codebook structure of the proposed method, ACBs use same open loop delays, a similar method for computing the delay contour, and FCB is a FPC codebook. The performance comparisons are shown in Table 2. Here, the proposed method uses $b_2 = 1.1$ and $\lambda_{th} = 1.25$ for the selection of codebook configuration.

Table 3 shows the performance of the proposed method with different codebook configuration selection methods.

Huffman Code	Codebook Configuration	Huffman Bits	ACB Bits	FCB Bits	Total bits
0	1 – 1 – 1	1	0	93	94
100	1 – 1 – 2	3	2	89	94
101	1 – 2 – 1	3	2	89	94
110	2 – 1 – 1	3	2	89	94
11100	1 – 2 – 1	5	4	85	94
11101	2 – 1 – 1	5	4	85	94
11110	2 – 2 – 1	5	4	85	94
11111	2 – 2 – 2	5	6	81	92

Table 1: Bits allocation for ACB/FCB Configuration over multiple subframes

Codebook Structure	Frame ACB Bits	Pulses in FCB	Frame FCB Bits	Total Bits	$WSNR_{Tot}$ / $WSNR_{Avg}$ (dB)
Structure-1	0	6	93	93	8.96 / 7.14
Structure-2	7	5	81	88	8.70 / 6.88
Structure-3	7	6	93	100	9.26 / 7.42
Proposed Method	0-6	5/6	81-93	94	9.12 / 7.27

Table 2: Performance comparison of the dynamic bit allocation method to fixed bit allocation methods

Bias Factor (b_2)	Prediction Gain Threshold (λ_{th})	Total WSNR (dB)	Average WSNR (dB)
1.0	0	9.009	7.140
1.1	0	9.127	7.269
1.1	1.25	9.123	7.273
1.0	1.25	9.094	7.220
Ideal Select		9.189	7.333

Table 3: Performance of the dynamic bit allocation method with respect to different bias factor and prediction gain threshhold used in the selection of codebook configuration

4.1. Discussion of Results

Table-2 shows that Structure-3 is 0.3 dB better than Structure-1 in terms of WSNR despite the fact that RCELP [5] was used for speech modification. The RCELP modifies the input speech based on the open loop delay and thereby enables the low distortion coding of the modified speech without the need of a delay adjustment factor. This gain of 0.3 dB (in Table-2) suggests that the closed loop delay adjust is useful even though speech has undergone the RCELP modification. However, Structure-3 uses 7 bits per frame more than Structure-1. The performance of Structure-2 being inferior to Structure-1 suggests that it is better overall to use 6 pulses in the FCB with no closed loop bits, and that RCELP modification is generally performing well.

The proposed method used only 1 extra bit compared to Structure-1 and its performance is half way between Structure-1 and Structure-3. If we use an "Ideal Select" (Table-3), i.e., the decision on which configuration to select is made after both ACB and FCB of both the configuration has been completed and their weighted distortions has been computed, then the proposed method is only 0.07 dB worse than Structure-3. From Table-3, it can be observed that if no bias factor (b_2= 1), and no gain threshold (λ_{th}) is used during selection of the codebook configuration then the performance gain when compared to Structure-1 is marginal. Thus, some kind of bias is necessary for the selection of the codebook configuration.

The codebook configuration selection with b_2=1.1, λ_{th}=1.25 results in around 63% of frames having no subframes using Configuration 2, 29% of frames having only one subframe using Configuration 2, 7% frames with two subframes having Configuration 2, and only 1% of frames with all subframes using Configuration 2. On the other hand, the Ideal Select approach has 53%, 35%, 10%, and 2% frames, having no subframes, only one subframe, two subframes, and all subframes using Configuration 2, respectively. Differences in the statistics of the Ideal Select and the selection method described in Section 2.1 and the difference in their WSNR performance suggest that some other parameters besides the one used in Section 2.1 may further improve the performance of the proposed dynamic bit allocation technique.

5. CONCLUSIONS

A method for dynamically allocating bits for the adaptive and fixed codebook of an analysis-by-synthesis speech codec based on the closed loop weighted mean squared error has been proposed. In this approach the bit allocation decision is made in each subframe. The weighted signal-to-noise-ratio shows that the dynamic bit allocation having two set of excitation codebooks is better than having a single set of excitation codebook. The technique is a part of EVRC-WB speech coding standard.

6. REFERENCES

[1] "Enhanced Variable Rate Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems," Document 3GPP2 C.P0014-C, Version 0.4, Sept. 2006.

[2] K. Ozawa, "4 kb/s multi-pulse based CELP speech coding using excitation switching,", *IEEE Proceedings of ICASSP*, pp. 189-192, March 1999.

[3] A. Ubale, A. Gersho, "A low-delay wideband speech codec at 24 kbps," *IEEE Proceedings of ICASSP*," pp. 165-168, May 1998.

[4] J. P. Ashley, E.M. Cruz-Zeno, U. Mittal, W. Peng, "Wideband coding of speech using a scalable pulse codebook," *IEEE Workshop on Speech Coding*, pp.148-150, Sept. 2000.

[5] W. B. Kleijn, P. Kroon, L. Cellario, D. Sereno, "A 5.85 kbits CELP algorithm for cellular applications," *IEEE Proceedings of ICASSP*, pp. 596-599, April 1993.