EMBEDDED CELP WITH ADAPTIVE CODEBOOKS IN ENHANCEMENT LAYERS AND MULTI-LAYER GAIN OPTIMIZATION

Jacek Stachurski

Texas Instruments, DSP Solutions R&D Center, Dallas TX

ABSTRACT

The paper describes an embedded CELP coder in which an adaptive codebook is included in every enhancement layer and the lower-layer codebook gains are re-optimized in the higher layers to further improve speech quality. Each layer maintains its own filter memories to generate required target vectors, adds adaptive and fixed codebook contributions, and re-optimizes all codebook gains to improve coder performance (multi-layer gain optimization). The common elements across the embedded layers include the lower-layer adaptive and fixed codebook entries. The pitch-lag used in the core layer is also re-used in the enhancement layers to maintain time synchronization between layers. Estimation and encoding of selected lower-layer parameters may take into account their estimated impact on the higher layers. The described Embedded CELP coder has been implemented in the Embedded Variable Bit-Rate (EV-VBR) codec standardized by ITU-T as Recommendation G.718. The characterization test results of the G.718 Embedded CELP are summarized.

Index Terms— Speech Coding, Embedded Coding, Scalable Coding, CELP

1. INTRODUCTION

Embedded coding enables bit-stream truncation at the decoder or any component of a telecommunication system allowing instantaneously bit-rate adjustment without the need for out-of-band signalling. This desired feature is the main design constraint of the ITU-T Recommendation G.718 approved in June 2008. The G.718 algorithm [4] is based on a two-stage coding structure: the lower two layers (R1 at 8 kb/s and R2 at 12 kb/s) are based on Code-Excited Linear Prediction (CELP). The higher three layers (R3 at 16 kb/s, R4 at 24 kb/s, and R5 at 32 kb/s) encode the weighted error signal from the lower layers using overlap-add Modified Discrete Cosine Transform (MDCT).

In a typical implementation of an embedded CELP coder [1, 2, 3], the R1 core layer Linear Prediction (LP) excitation target is generated in the perceptually-weighted domain, the adaptive and fixed codebook contributions are chosen, and the codebook gains are quantized. The adaptive-codebook contribution is included only in the core layer with the enhancement layers providing additional fixed codebooks (Fig. 1). The fixed-codebook structure of the enhancement layers can be modified to account for the characteristics of the signal generated in the lower layers [2], but the adaptive codebook is not used. The parameters encoded in each layer are re-used in the higher layers. The R2 target excitation is computed from the R1 target by subtracting the R1 codebook contributions. The target vector is updated in consecutive layers by subtracting the additional fixed-codebook contributions.

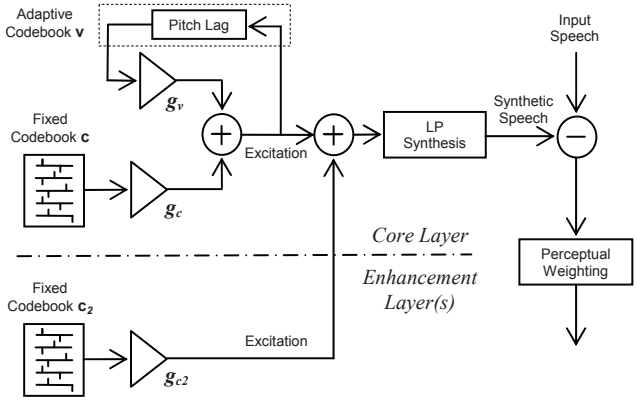


Figure 1: Embedded CELP with fixed codebooks in enhancement layers

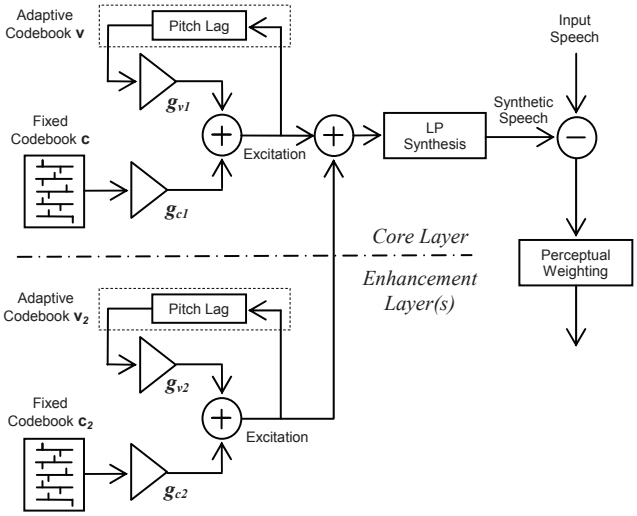


Figure 2: Embedded CELP with adaptive and fixed codebooks in enhancement layers. The lower-layer gains are re-optimized when used with the higher layers

We propose using both, adaptive codebook and fixed-codebook contributions in all layers of an embedded CELP coder (Fig. 2). In this configuration, each layer of the encoder optimizes the LP excitation parameters with respect to the target vector generated for that layer. Every layer maintains its own filter memories to generate required target vectors, adds adaptive and fixed codebook contributions, and re-optimizes the lower-layer

codebook gains for improved performance (multi-layer gain optimization). Common elements across embedded layers include the lower-layer adaptive and fixed codebook entries. The pitch-lag used in the core layer is re-used in the enhancement layers to maintain time synchronization between layers. Estimation and encoding of selected lower-layer parameters may also take into account their estimated impact on the higher layers.

Section 2 provides an outline of the proposed Embedded CELP design and discusses various coder configurations. Section 3 presents the Embedded CELP implemented in the ITU-T Recommendation G.718 and summarizes relevant subjective test results.

2. EMBEDDED CELP

The goal of the proposed Embedded CELP design is to enhance speech quality by effectively including an adaptive codebook contribution in every layer and, as far as possible within specified bit rates, re-optimize the lower-layer LP excitation parameters for improved performance with the higher layers. In practice, there are typically no spare bits available in the enhancement layers to modify the adaptive and fixed codebook entries that are selected in the lower layers. However, it is possible with relatively few bits to update the gains corresponding to these codebooks so that they provide an additional positive impact when combined with the higher enhancement-layer codebooks. It is also possible to select lower-layer codebook entries so that they not only provide high performance in lower layers, but also result in speech quality improvement when used in conjunction with higher layers.

In the next sections, the following notation is used:

$\mathbf{u}$, $\hat{\mathbf{u}}$: target and coded excitation vectors
$\mathbf{v}$, $\mathbf{c}$: adaptive and fixed (e.g. algebraic) codebook entries
$\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$: vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{c}$ in perceptually weighted domain
g_v, $\hat{g}_v$, g_c, $\hat{g}_c$: optimal and encoded gains for $\mathbf{v}$ and $\mathbf{c}$

2.1. Core Layer

In the R1 layer, the CELP decoder generates the LP excitation $\hat{\mathbf{u}}$,

$$\hat{\mathbf{u}} = \hat{g}_v \mathbf{v} + \hat{g}_c \mathbf{c} .$$

At the encoder, the following simplified steps are executed to encode the LP excitation $\hat{\mathbf{u}}$:

- Target vector $\mathbf{u}$ computation
- Adaptive codebook $\mathbf{v}$ search (pitch-lag estimation, typically performed in two stages: open-loop and closed-loop search),

$$\arg\min_{\mathbf{v}} (\mathbf{u} - g_v \mathbf{v})^2$$

- Fixed codebook $\mathbf{c}$ search,

$$\arg\min_{\mathbf{c}} (\mathbf{u} - g_v \mathbf{v} - g_c \mathbf{c})^2$$

- With $\mathbf{c}$ selected, closed-loop search for codebook gains,

$$\arg\min_{\hat{g}_v, \hat{g}_c} (\mathbf{u} - \hat{g}_v \mathbf{v} - \hat{g}_c \mathbf{c})^2$$

Note that minimization of all above errors is performed in perceptually-weighted domain using filtered vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ that correspond to vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{c}$.

2.2. Enhancement Layer

In the R2 enhancement layer, the Embedded CELP decoder is set to generate the decoded excitation $\hat{\mathbf{u}}_2$ as

$$\hat{\mathbf{u}}_2 = \hat{g}_{v1} \mathbf{v}_1 + \hat{g}_{c1} \mathbf{c}_1 + \hat{g}_{v2} \mathbf{v}_2 + \hat{g}_{c2} \mathbf{c}_2 ,$$

with subscript $_1$ indicating parameters related to the R1 core layer and subscript $_2$ referring to parameters introduced in the R2 enhancement layer. As stated earlier, there are typically no spare bits available to modify the adaptive and fixed codebook entries selected in R1 so in practice $\mathbf{v}_1 = \mathbf{v}$ and $\mathbf{c}_1 = \mathbf{c}$ which leads to

$$\hat{\mathbf{u}}_2 = \hat{g}_{v1} \mathbf{v} + \hat{g}_{c1} \mathbf{c} + \hat{g}_{v2} \mathbf{v}_2 + \hat{g}_{c2} \mathbf{c}_2 .$$

Note that $\hat{g}_{v1}$ and $\hat{g}_{c1}$ (the quantized gains applied to codevectors $\mathbf{v}$ and $\mathbf{c}$ when decoding $\hat{\mathbf{u}}_2$ in R2) are different from $\hat{g}_v$ and $\hat{g}_c$ (the gains applied to $\mathbf{v}$ and $\mathbf{c}$ when decoding the R1 core-layer excitation $\hat{\mathbf{u}}$). While it may be impractical to modify vectors $\mathbf{v}$ and $\mathbf{c}$, modifying $\hat{g}_v$ and $\hat{g}_c$ is feasible with only a small number of additional bits (e.g., as described in Section 3).

Given the R2 decoder, the following steps may be executed to encode the excitation $\hat{\mathbf{u}}_2$:

- Target vector $\mathbf{u}_2$ computation with analysis-by-synthesis procedure performed based on the past R2 excitation (note that $\mathbf{u}_2$ is therefore different from the R1 target $\mathbf{u}$)
- Adaptive codebook $\mathbf{v}_2$ generation. To maintain time synchrony between layers and to save bits, the same pitch-lag is used as in the R1 core layer. To generate $\mathbf{v}_2$, the pitch-lag is applied to the past R2 enhancement component $\hat{\mathbf{u}}_{22}$,

$$\hat{\mathbf{u}}_{22} = \hat{g}_{v2} \mathbf{v}_2 + \hat{g}_{c2} \mathbf{c}_2$$

- Fixed codebook $\mathbf{c}_2$ search,

$$\arg\min_{\mathbf{c}_2} (\mathbf{u}_2 - g_{v1} \mathbf{v} - g_{v2} \mathbf{v}_2 - g_{c1} \mathbf{c} - g_{c2} \mathbf{c}_2)^2$$

- With $\mathbf{c}_2$ selected, closed-loop search for codebook gains,

$$\arg\min_{\hat{g}_{v1}, \hat{g}_{v2}, \hat{g}_{c1}, \hat{g}_{c2}} (\mathbf{u}_2 - \hat{g}_{v1} \mathbf{v} - \hat{g}_{v2} \mathbf{v}_2 - \hat{g}_{c1} \mathbf{c} - \hat{g}_{c2} \mathbf{c}_2)^2$$

As stated before, the minimization of all errors is performed in perceptually-weighted domain.

The above configuration of the R2 enhancement layer was verified in the qualification phase of the G.718 development and judged to provide clear improvement over a typical embedded CELP with only a fixed-codebook in the enhancement layer. Extensive objective evaluations were first performed with the Perceptual Evaluation of Speech Quality (PESQ) measure. The PESQ scores increase as high as 0.110 in some cases and by about 0.060 on average over a range of narrowband and wideband test conditions. The technology was successfully validated within the G.718 EV_VBR coder candidate in formal subjective tests.

Note that variants of the above processing steps may be implemented. For example, to reduce complexity, the $\mathbf{c}_2$ fixed codebook search may be performed based on the target updated with the quantized gains from the R1 core layer $\hat{g}_v$ and $\hat{g}_c$:

$$\arg\min_{\mathbf{c}_2} (\mathbf{u}_2 - \hat{g}_v \mathbf{v} - g_{v2} \mathbf{v}_2 - \hat{g}_c \mathbf{c} - g_{c2} \mathbf{c}_2)^2 .$$

This approach reduces complexity because the optimal gains g_{v1} and g_{c1} do not need to be calculated in this case.

In the context of the stated R2 enhancement-layer decoder, we observe that a traditional embedded CELP employs a simplified version of the configuration outlined above:

$$\hat{\mathbf{u}}_2 = \hat{g}_v \mathbf{v} + \hat{g}_c \mathbf{c} + \hat{g}_{c2} \mathbf{c}_2 .$$

The $\hat{g}_v$ and $\hat{g}_c$ gains are re-used in R2 without modification and the adaptive-codebook contribution $\mathbf{v}_2$ is not present. The R2 target vector is also re-used from the R1 core layer, $\mathbf{u}_2 = \mathbf{u}$. The fixed codebook search is performed minimizing

$$\arg\min_{\mathbf{c}_2} \left(\mathbf{u}'_2 - g_{c2}\mathbf{c}_2\right)^2 ,$$

where $\mathbf{u}'_2 = \mathbf{u} - \hat{g}_v \mathbf{v} - \hat{g}_c \mathbf{c}$. The $\mathbf{c}_2$ codebook gain g_{c2} is then quantized with

$$\arg\min_{\hat{g}_{c2}} \left(\mathbf{u}'_2 - \hat{g}_{c2}\mathbf{c}_2\right)^2 .$$

Again, the minimization of all errors is performed in perceptually-weighted domain.

2.3. Higher Enhancement Layers

In the R3 enhancement layer, the Embedded CELP decoder would generate the excitation $\hat{\mathbf{u}}_3$ as

$$\hat{\mathbf{u}}_3 = \hat{g}_{v13}\mathbf{v} + \hat{g}_{c13}\mathbf{c} + \hat{g}_{v23}\mathbf{v}_2 + \hat{g}_{c23}\mathbf{c}_3 + \hat{g}_{v3}\mathbf{v}_3 + \hat{g}_{c3}\mathbf{c}_3 ,$$

where $\hat{g}_{v13}$, $\hat{g}_{c13}$, $\hat{g}_{v23}$, $\hat{g}_{c23}$ are the R1 and R2 layer codebook gains re-optimized for usage in R3, and $\hat{g}_{v3}$, $\mathbf{v}_3$, $\hat{g}_{c3}$, $\mathbf{c}_3$ are the additional LP excitation parameters introduced in the new layer. With each enhancement layer, the number of gains to be re-optimized and encoded increases leading to increased bit rate and complexity. Depending on the number of implemented layers, various Embedded CELP decoder configurations may be devised to reduce complexity by reducing the number of gains to be re-optimized. In one investigated variant, scaling factors s are introduced with consecutive layers decoded as

R1: $\hat{\mathbf{u}} = \hat{g}_v \mathbf{v} + \hat{g}_c \mathbf{c}$,

R2: $\hat{\mathbf{u}}_2 = \hat{g}_{v2}\left(\mathbf{v} + \mathbf{v}_2\right) + \hat{g}_{c2}\left(s_{12}\mathbf{c} + \mathbf{c}_2\right)$,

R3: $\hat{\mathbf{u}}_3 = \hat{g}_{v3}\left(\mathbf{v} + \mathbf{v}_2 + \mathbf{v}_3\right) + \hat{g}_{c3}\left(s_{13}\mathbf{c} + s_{23}\mathbf{c}_2 + \mathbf{c}_3\right)$

etc., where s_{12}, s_{23}, s_{23} are the scaling factors applied to the lower-layer fixed-codebook entries when they are used with the higher enhancement layers. In this configuration, only two codebook gains are optimized and encoded in each layer, e.g. $\hat{g}_{v3}$ and $\hat{g}_{c3}$ in R3, with the scaling factors allowing application of variable gains to codebooks used in different layers. For bit-rate efficiency, differential gain encoding with respect to lower-layer gains can be employed. The scaling factors may be fixed or adaptive; in a system with three enhancement layers, we obtained good results with the following fixed scaling factors:

R2: $s_{12} = 1.375$,

R3: $s_{13} = 1.75$, $s_{23} = 1.375$,

R4: $s_{14} = 2.125$, $s_{24} = 1.75$, $s_{34} = 1.375$.

In general, the fixed-codebook contributions from the lower layers are assigned larger weights in the LP excitation mix since they are expected to capture more relevant features of the target excitation.

Given the example decoder, the R3 encoder would execute the following steps:

- Target vector $\mathbf{u}_3$ computation with analysis-by-synthesis procedure performed based on the past R3 excitation
- Adaptive codebook $\mathbf{v}_3$ generation with the R1 pitch-lag applied to the past R3 enhancement-layer excitation component $\hat{\mathbf{u}}_{33}$,

  $$\hat{\mathbf{u}}_{33} = \hat{g}_{v3}\mathbf{v}_3 + \hat{g}_{c3}\mathbf{c}_3$$

- Fixed codebook $\mathbf{c}_3$ search,

  $$\arg\min_{\mathbf{c}_3} \left[\mathbf{u}_3 - g_{v3}\mathbf{v}'_3 - g_{c3}\left(s_{13}\mathbf{c} + s_{23}\mathbf{c}_2 + \mathbf{c}_3\right)\right]^2 ,$$

  where $\mathbf{v}'_3 = \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3$

- With $\mathbf{c}_3$ selected, closed-loop search for codebook gains,

  $$\arg\min_{\hat{g}_{c3}, \hat{g}_{c3}} \left(\mathbf{u}_3 - \hat{g}_{v3}\mathbf{v}'_3 - \hat{g}_{c3}\mathbf{c}'_3\right)^2 ,$$

  where $\mathbf{c}'_3 = s_{13}\mathbf{c} + s_{23}\mathbf{c}_2 + \mathbf{c}_3$

This simplified Embedded CELP implementation can be advantageous when three or more enhancement layers are considered, but it is not preferred when only one enhancement layer is needed.

2.4. Cross-Layer parameter optimization

With distinctive target vectors generated for every embedded layer, the lower-layer parameters may be selected to provide an improved performance in the higher layers. We experimented with lower-layer fixed-codebook selection based on higher-layer targets and obtained improved performance in the higher layers at the cost of slight performance drop in the lower-layers (the effect becomes more pronounced as the number of enhancement layers increases). This technique would be useful in a system in which the major coder operation mode is at its maximum bit rate, with lower bit-rates used less often.

The pitch-lag estimation may also be performed with respect to the higher-layer target vectors, not the R1 core-layer target. In the G.718 narrowband mode, we implemented a variant of this approach with the closed-loop pitch-lag search based on an LP excitation generated with the four R1 and R2 codebook contributions scaled with optimal gains [5]. This is done to make the closed-loop pitch search independent of gain-quantization errors and to provide improved pitch estimation across layers.

3. EMBEDDED CELP IN G.718

In June 2008, ITU-T SG16 approved a new Embedded Variable Bit-Rate (EV-VBR) coder as Recommendation G.718. The G.718 EV-VBR codec is an 8 to 32 kb/s embedded coder comprising five layers with the lower two layers, R1 at 8 kb/s and R2 at 12 kb/s, based on CELP technology. The R1 and R2 layers include several coding modes optimized for different input signals; the Generic and Voiced modes of the R2 enhancement layer are encoded with the Embedded CELP as outlined in Section 2.2. The error from the CELP layers is further encoded by three higher layers, R3 at 16 kb/s, R4 at 24 kb/s, and R5 at 32 kb/s, in a transform domain

using overlap-add MDCT. In addition to scalable design, G.718 is highly robust to frame erasures enhancing speech quality when used in IP transport applications. The encoder accepts wideband (WB) and narrowband (NB) signals sampled at 16 kHz and 8 kHz, respectively. The decoder output can also be WB or NB; while WB rendering is provided for all layers, NB rendering is implemented only for the two CELP layers, R1 and R2. The input signal is processed using 20 ms frames. Independently of the input sampling rate, the CELP internal sampling frequency is 12.8 kHz.

The computation of the R2 target vectors, the R2 adaptive codebook generation, the optimal gain calculation for the R1 fixed and adaptive codebooks applied in the R2 layer, the optimal gain calculation for the R2 adaptive-codebook, and the R2 fixed codebook search are described in detail in [5]. The four R2 gains $g_{v1}, g_{v2}, g_{c1}, g_{c2}$ are jointly quantized with respect to the gains encoded in R1, $\hat{g}_v$ and $\hat{g}_c$. Four bits are used for the gain quantization in each 5 ms sub-frame. The four-dimensional codebook entries specify gain correction factors $\gamma_{v1}, \gamma_{v2}, \gamma_{c1}, \gamma_{c2}$ to be applied to the coded R1 gains: $\hat{g}_{v1} = \gamma_{v1}\hat{g}_v$, $\hat{g}_{v2} = \gamma_{v2}\hat{g}_v$, $\hat{g}_{c1} = \gamma_{c1}\hat{g}_c$, $\hat{g}_{c2} = \gamma_{c2}\hat{g}_c$. The gain corrections that minimize the perceptually weighted mean-square error are chosen with

$$\arg\min_{\gamma_{v1},\gamma_{v2},\gamma_{c1},\gamma_{c2}} \left[\mathbf{x}_2 - \hat{g}_v\left(\gamma_{v1}\mathbf{y} + \gamma_{v2}\mathbf{y}_2\right) - \hat{g}_c\left(\gamma_{c1}\mathbf{z} + \gamma_{c2}\mathbf{z}_2\right)\right]^2,$$

where $\mathbf{x}_2, \mathbf{y}, \mathbf{y}_2, \mathbf{z}, \mathbf{z}_2$ represent perceptually-weighted vectors.

The G.718 codec performance was formally evaluated in ITU-T Characterization Tests with subjective Mean Opinion Score (MOS) experiments. For a general R1 and R2 Embedded CELP performance overview, the average scores for NB and WB clean speech and background noise conditions are summarized in Tables 1 and 2. The NB clean speech results are averaged over the three tested input levels: -16 dBov, -26 dBov, and -36 dBov. The WB clean speech R2 was tested for the -26 dBov nominal speech level only. The average scores for the NB background noise conditions include 15 dB car noise, 20 dB street noise, 20 dB office noise, 25 dB babble noise, 15 dB interfering talker, and 25 dB background music. The average results for WB noise conditions include 20 dB street noise, 20 dB office noise, and 25 dB babble noise. All results are averaged over two listening laboratories that tested each condition. The clean speech tests were conducted using Absolute Category Rating (ACR) and the noise condition tests were performed using Degradation Category Rating (DCR) which accounts for different absolute scores in the two test sets. The G.718 codec provides an overall performance improvement over reference coders at comparable rates: NB and WB G.718 R1 at 8 kb/s and R2 at 12 kb/s vs. NB G.729 at 8 kb/s, NB G.729E at 11.8 kb/s, and WB G.722.2 at 8.85 kb/s and 12.65 kb/s. In the NB clean speech and background noise conditions and the WB clean speech, the G.718 R2 scores approach those of the "Direct" anchor. For the WB background noise case where saturation does not affect the results, the R2 performance improvement over R1 is significant. Detailed results for all conditions can be found in [6].

Table 1: Average narrowband MOS results

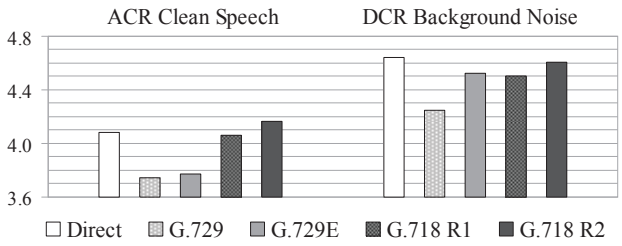
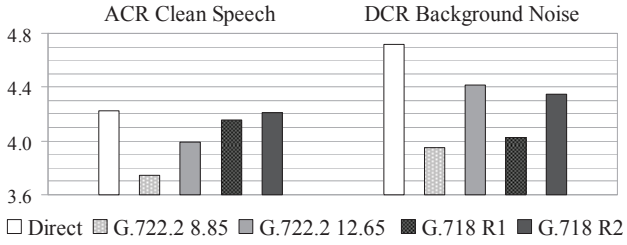


Table 2: Average wideband MOS results



4. CONCLUSIONS

We presented an embedded CELP framework with adaptive codebooks in enhancement layers and multi-layer gain optimization. This coding approach was successfully implemented in the ITU-T Recommendation G.718. The verification tests conducted in the qualification phase of the G.718 development showed clear preference for this technology over a traditional embedded CELP and the tests performed in the characterization phase confirm very good performance of this embedded design.

5. ACKNOLEGMENTS

We would like to thank a number of colleagues who participated in the Embedded CELP verification tests and provided valuable feedback in the course of its integration into the G.718 coder. In particular, we would like to thank Erdem Ertan, Vishu Viswanathan, Tommy Vaillancourt, Milan Jelinek, Jon Gibbs, Jonas Svedberg, Anssi Rämö, and all other members of the G.718 consortium.

5. REFERENCES

[1] R. D. De Iacovo and D. Sereno, "Embedded CELP Coding for Variable Bit-Rate Between 6.4 and 9.6 kbit/s," in *Proc. IEEE ICASSP*, Toronto, pp. 681-684, May 1991

[2] T. Nomura et al., "A Bitrate and Bandwidth Scalable CELP Coder," in *Proc. IEEE ICASSP*, Seattle, pp. 341-344, May 1998

[3] S. Ragot et al., "ITU-T G.729.1 an 8-32 Kbit/S Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice Over IP," in *Proc. IEEE ICASSP*, Honolulu, pp.529-532, April 2007

[4] Tommy Vaillancourt et al., "ITU-T EV-VBR: a robust 8-32 kbit/s scalable coder for error prone telecommunications channels," *Proc. EUSIPCO*, Lausanne, August 2008

[5] ITU-T Recommendation G.718, "Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s," June 2008

[6] ITU-T Q7/SG12 Technical Contribution AH-08-44, "Summary of results for G.EV-VBR," Lannion, April 2008