ITU-T G.729 EXTENSION AT 6.4 KBPS

E. Ekudden∗, R. Hagen∗, B. Johansson∗, S. Hayashi†, A. Kataoka†, S. Kurihara†

* Audio and Visual Technology Research, Ericsson Radio Systems AB, S-164 80 Stockholm, Sweden
† NTT Human Interface Labs. 3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper describes the 6.4 kbit/s CS-ACELP coder being standardized as annex D to ITU-T G.729. The coder is based on the same building blocks as the 8 kbit/s G.729 to facilitate low complexity extensions to G.729 in terms of additional memory usage. It is fully switchable with the 8 kbit/s coder and provides additional flexibility to existing and emerging G.729 applications. The fixed codebook is a 2-pulse algebraic codebook. The adaptive codebook quantization has been changed and a new conjugate structure gain codebook is used. In order to compensate for the sparser algebraic codebook, an adaptive post-processing technique is used to enhance the quality for unvoiced speech and background noise sounds. Subjective tests have indicated that the coder has a performance close to that of G.729, and equivalent to that of G.723.1 at 6.3 kbit/s for speech.

1. INTRODUCTION

Recent speech coding standardization activities in ITU-T have led to the adoption of the 8 kbit/s G.729 Conjugate Structure Algebraic CELP (CS-ACELP) [1]. The G.729 coder has a 10 ms frame size and 5 ms look-ahead, leading to a moderate delay. G.729 provides "toll quality" for speech and is expected to be deployed in applications requiring high speech quality, low bit-rate and at the same time medium delay.

Following the completion of the standardization of the 8 kbit/s coder in 1996, a new effort to standardize extensions to the basic algorithm operating at 6.4 kbit/s and approximately 12 kbit/s was started. This will increase the flexibility and applicability of G.729. The requirements for the lower bit-rate extension were set mainly in terms of the quality of G.726 ADPCM at 24 kbit/s and G.729 itself. An additional constraint to only allow 10% increase in memory usage was set to mandate reuse of most of the basic CS-ACELP algorithm in order to provide easy addition of the lower bit-rate mode to existing 8 kbit/s implementations.

In September 1997, Ericsson and NTT submitted 6.4 kbit/s candidates to ITU-T. The subjective qualification test results indicated that the two coders had similar overall performance. Also the structure of the proposals was similar. The differences were primarily in the structure of the gain codebook and the algebraic codebook. Between September and January, an evaluation and optimization phase took place, where a single coder was developed using parts from both proposals. The final coder was shown to have improved performance overall. It was presented to ITU-T in January 1998, and approved (determined) as annex D to G.729.

2. REQUIREMENTS

The requirements were set in terms of subjective quality, computational complexity, RAM, and ROM usage. The design constraints are summarized in Table I. The main purpose of the strict design constraints was to allow cost efficient updates to 8 kbit/s implementations. In addition, the delay should be lower or equal to that of G.729, i.e. a maximum algorithmic delay of 15 ms was allowed.

Table I. Design constraints.

Parameter	Requirement
Complexity	≤ G.729
RAM	≤ 10% increase for G.729 implementation
ROM	≤ 10% increase for G.729 implementation

The requirements in terms of subjective quality are summarized in Table II. These were set mainly in terms of the quality of G.726 ADPCM at 24 kbps (G.726-24), which is a lower rate extension to 32 kbit/s ADPCM. The objectives were set mainly in terms of the quality of G.729 itself. For detected frame erasures (FER), the requirements were set in terms of increase in number of Poor or Worse ("PoW") votes.

Table II. Main speech quality requirements.

Condition	Requirement	Objective
Error free	> G.726-24	G.729
High/Low level	G.726-24	G.729
BER 10^{-3}	G.726–24, BER 10^{-3}	G.729, BER 10^{-3}
3% Random FER	G.729 + 10% PoW	G.729+5% PoW
3% Bursty FER	G.729 + 10% PoW	G.729+5% PoW
Tandem	G.726-24 tandem	f.f.s
Car noise	G.726-24	f.f.s
Babble noise	G.726-24	f.f.s
Interfering talker	G.726-24	f.f.s

The coder should be able to switch between 8 and 6.4 kbit/s with a quality equivalent to G.726 when switching between 24 and 32 kbit/s.

3. CODER DESCRIPTION

In this section, the 6.4 kbit/s CS-ACELP, G.729 Annex D speech coding algorithm is described. The description is organized in an overview followed by descriptions of each building block. The differences to the 8 kbit/s main body G.729 are pointed out for each building block. For more details, refer to the description of the 8kbit/s G.729 standard [1,2].

3.1. Overview

Figure 1 illustrates the principle of the encoding algorithm. It follows the Linear Prediction Analysis-by-Synthesis (LPAS) principle [3]. The coder operates with a frame size of 10 ms and two 5 ms subframes. In addition, the linear prediction (LP) analysis uses a look-ahead of 5 ms, resulting in a total algorithmic delay of 15 ms. The main building blocks are LP analysis and quantization for the short-term spectral envelope, an adaptive codebook for long term (pitch) prediction, an algebraic codebook for innovation coding, and a conjugate structure vector quantizer for gain quantization. Table II shows the bit allocation between these blocks for the two subframes as well as the total per 10 ms frame.

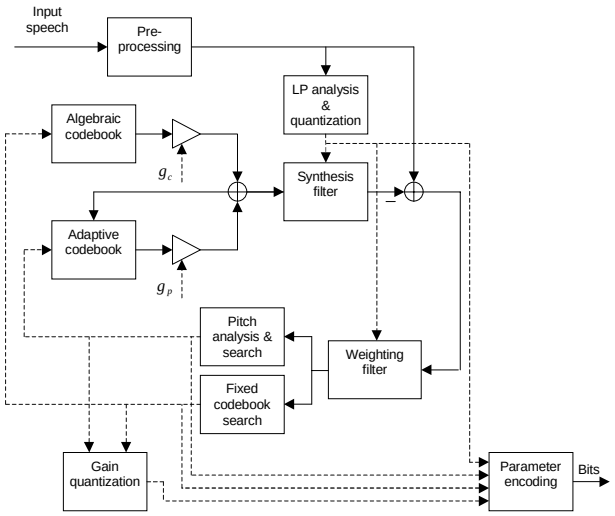

Figure 1. Principle of encoder.

Table III. Bit allocation for the 6.4 kb/s CS-ACELP algorithm.

Parameter	Subframe 1	Subframe 2	Total
LP coefficients			18
Adaptive CB	8	4	12
Algebraic CB index	9	9	18
Algebraic CB sign	2	2	4
CB gains (stage 1)	3	3	6
CB gains (stage 2)	3	3	6
Total			64

Figure 2 illustrates the principle of the decoder. It includes anti-sparseness processing for the algebraic codebook and post-processing of the synthesized speech signal in addition to the building block used in the encoder.

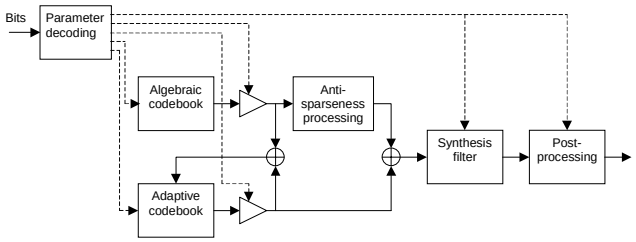

Figure 2. Principle of decoder.

3.2. Pre-processing

Two pre-processing functions are applied to the input signal: 1) Signal downscaling by dividing the input by 2, and 2) High-pass filtering with a second order pole/zero filter with cut-off frequency 140 Hz. The pre-processing is exactly as in the main body G.729.

3.3. LP analysis and quantization

A 10th order Liner Prediction (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 are applied to the autocorrelation function.

The resulting LP coefficients are converted to Line Spectrum Frequencies (LSFs) prior to quantization. A switched 4th order MA prediction requiring one bit is used to predict the LSFs of the current frame. The prediction residual is quantized using a 2-stage VQ. The first stage is a 7-bit VQ for all 10 dimensions. The second stage consists of a 2-split VQ with two 5-dimensional, 5-bit VQs.

The LP synthesis filter is given by:

$$H(z) = \frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{i=1}^{10} \hat{a}_i z^{-i}} \qquad (1)$$

where a_i are the quantized LP coefficients.

The LP analysis and quantization is identical to the main body G.729.

3.4. Perceptual weighting filter

The perceptual weighting filter is computed from the unquantized LP coefficients a_i by:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} . \qquad (2)$$

The factors γ_1 and γ_2 are adapted to the spectral shape of the input signal. If the input signal is characterized as flat, the values 0.94 and 0.6 are used. Otherwise, γ_1 is set to 0.98 and γ_2 is a function of the strength of the resonances in the LP synthesis filter so that the stronger the main resonance, the higher the value (it is bounded between 0.4 and 0.7).

The perceptual weighting filter is identical to the main body of G.729.

3.5. Adaptive codebook

Once 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 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. The closed-loop search is performed as:

1. Search 6 integer delays around the open-loop estimate T_o to find the best integer delay T_1.

2. If T_1 is less than 85, search the fractional values around T_1.

In the 2nd subframe, the adaptive codebook uses a 4-bit delta-coded pitch delay. The delay is coded relative to the pitch delay of the 1st subframe rounded to integer resolution. The search is performed as:

1. Search 10 integer delays around the integer delay of the 1st subframe to find the best integer delay T_2.
2. If T_2 is one of the 2 middle values of the integer search range, search the fractional values around T_2.

The open-loop pitch analysis and the closed-loop search in the 1st subframe are identical to the main body G.729. In the 2nd subframe, the main body uses 5-bit delays instead. The integer search range is the same. The additional delay values are obtained by having fractional resolution in the entire integer search range.

3.6. Fixed codebook

The fixed codebook employs the algebraic structure with two signed pulses in two overlapping tracks. Table IV shows the track table for the algebraic codebook. Each pulse has one-bit sign. The 1st pulse can take on one of 16 positions whereas the 2nd pulse is located at one of 32 positions. This gives 4 and 5 bits for position coding, a total of 9 position bits. The structure of the algebraic codebook is different compared to the main body G.729 which uses 4 signed pulses in 4 non-overlapping tracks.

The search procedure for the fixed codebook follows the algebraic codebook search used in the main body G.729 except that 2 pulses need to be searched instead of 4 pulses. Thereby there are only 2 inner search loops to test pulse positions instead of 4. These 2 loops perform exhaustive search of pulse positions. The sign of each pulse in each position is pre-set to the sign of the target signal. The efficient procedure for computation of necessary correlation terms in the main body G.729 is adopted with modifications to the new structure of the codebook. The search complexity for the algebraic codebook is significantly less than for the main body G.729.

Table IV. Track table for algebraic codebook.

Pulse	Sign	Position
i_o	±1	1,3,6,8,11,13,16,18,21,23,26,28,31,33,36,38
i_1	±1	0,1,2,4,5,6,7,9,10,11,12,14,15,16,17,19,20, 21,22,24,25,26,27,29,30,31,32,34,35,36,37, 39

3.7. Gain quantization

Gain quantization starts with a 4th order MA prediction of the fixed codebook gain. The prediction is performed in the mean-removed log-energy (in dB) domain. A 2-stage conjugate structure VQ [4] is used for quantization of the adaptive codebook gain and the prediction residual for the fixed codebook gain. Each stage uses 3 bits to give a total of 6 bits. The VQ codebook is trained with the condition of 0.1% bit error rate with random distribution.

The main body G.729 uses the same conjugate structure VQ technique but with 3 and 4 bits in the 1st and 2nd stage giving a total of 7 bits.

3.8. Post-processing

Anti-sparseness processing

Due to the sparse algebraic codebook with only 2 pulses per 40 samples subframe, a novel anti-sparseness processing [5] of the fixed codebook signal is performed. The fixed codebook vector is circularly convoluted with an impulse response with the properties:

1. Unit magnitude spectrum. Thus, the magnitude of the fixed codebook vector is left unaltered.
2. Semi-random high-frequency phase spectrum. Thus, a semi-random component is added to the high-frequency phase of the fixed codebook vector.

The annoying artifacts caused by the sparseness is 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. The adaptive codebook gain g_p and the fixed codebook gain are used to select one of three impulse responses with the following properties:

1. Strong modification: Random phase between $-\pi$ and π in the frequency range from 2 to 4 kHz.
2. Medium modification: Random phase between $-\pi/2$ and $\pi/2$ in the frequency range from 3 to 4 kHz.
3. No modification.

The impulse responses are adaptively selected according to the following procedure:

1. Select impulse response 1 if g_p <0.6, select impulse response 2 if g_p is in the range 0.6 to 0.9, select impulse response 3 if g_p >0.9.
2. Compute an onset indicator which 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_p and the previous 5 values. If the result is less than 0.6, select impulse response 1.
4. If onset is indicated and the impulse response is not 2, increment the impulse response selected by 1.

This adaption 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 adaption is based on the quantized gain values, no extra information is needed to select the correct impulse response.

The anti-sparseness processing does not exist in the main body G.729 algorithm.

Post-processing

The post-processing of the coded speech signal is identical to that of the main body G.729 and includes:

1) Long-term (pitch) postfiltering to enhance the pitch periodicity of voiced speech segments.

2) Short-term (formant) postfiltering to enhance the formant structure.

3) Tilt compensation to compensate for the tilt in the short-term postfilter.

4) Adaptive gain control to compensate for gain differences between the coded speech signal and the post-filtered signal.

5) High-pass filtering with a 2nd order pole/zero filter with a cut-off frequency of 100 Hz.

6) Signal upscaling by a factor of 2 to invert the down-scaling in the pre-processing.

4. PERFORMANCE EVALUATION

The quality has been extensively evaluated in several languages. Figure 3, 4, and 5 summarize results obtained from experiments conducted in japanese. Each experiment included 24 naïve listeners. Results are presented for the 6.4 kbit/s extension to G.729 (G.729 6.4k), G.726 at 24 kbit/s (G.726 24k), 8 kbit/s main body G.729 (G.729), and G.723.1 at 6.3 kbit/s (G.723.1 6.3k).

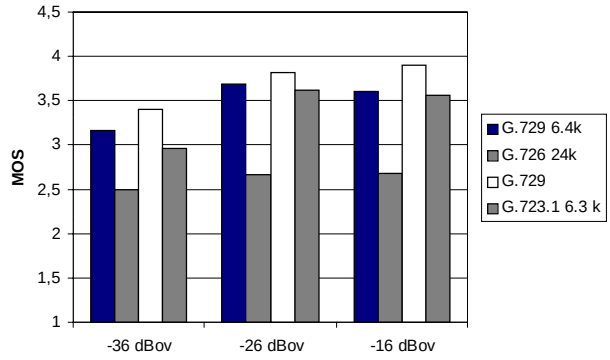

Figure 3. Subjective test results from ACR test with clean speech at the input levels -16 dB, -26 dB, and –36 dB.

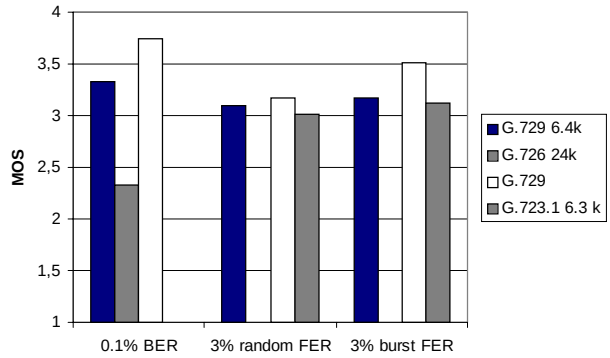

Figure 4. Subjective test results from ACR test with clean speech for the error conditions: 0.1% BER, 3% random FER, and 3% bursty FER.

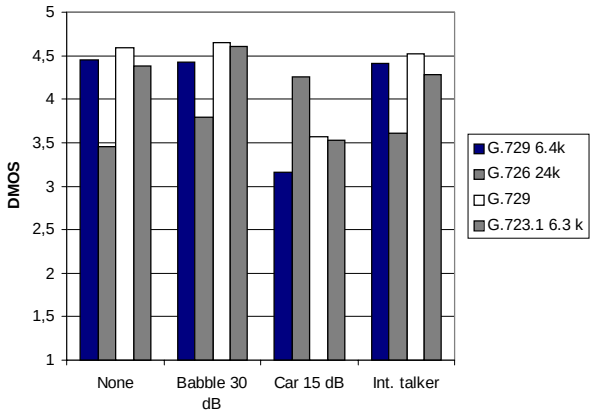

Figure 5. Subjective test results from DCR test with the background noise conditions: none, babble, car, and interfering talker.

The results can be summarized in the following way. For clean speech the quality is significantly higher than the requirement, G.726 at 24 kbit/s, and only slightly lower than G.729. It is equivalent to G.723.1 at 6.3 kbit/s. In background noise the quality is better than G.726 at 24 kbit/s, except for car noise, where the quality is lower, which is also the case for G.729.

5. CONCLUSIONS

The 6.4 kbit/s CS-ACELP extension to G.729 employs the basic structure of G.729 with new fixed, adaptive, and gain codebooks. Hence, the additional memory is in the order of only 10% which should allow efficient extensions to 8 kbit/s coder implementations. The novel phase-dispersion post-processing has made it possible to use only two pulses per subframe in the algebraic codebook. Under most conditions, the coder exceeds the requirements, providing high quality for bandwidth limited systems.

REFERENCES

[1] ITU-T Recommendation G.729, Coding of speech at 8 kbit/s using Conjugate Structure-Algebraic Code Excited Linear Prediction (CS-ACELP).

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

[4] A Kataoka, T. Moriya, and S. Hayashi, "An 8-kb/s Conjugate Structure CELP (CS-CELP) Speech Coder," IEEE Trans. on SAP, Vol. 4, No. 6, pp.401-411, Nov. 1996.

[5] 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.