DESCRIPTION OF ITU-T RECOMMENDATION G.729 ANNEX A: REDUCED COMPLEXITY 8 KBIT/S CS-ACELP CODEC

R. Salami *C. Laflamme* *B. Bessette* *J-P. Adoul*

Department of Electrical Engineering, University of Sherbrooke,
Sherbrooke, Québec, Canada, J1K 2R1

ABSTRACT

This paper describes the recently adopted ITU-T Recommendation G.729 Annex A (G.729A) for encoding speech signals at 8 kbit/s with low complexity. G.729A has been selected as the standard speech coding algorithm for multimedia digital simultaneous voice and data (DSVD). G.729A is bitstream interoperable with G.729; i.e., speech coded with G.729A can be decoded with G.729, and vice versa. As G.729, it uses the CS-ACELP algorithm with 10 ms frames. However, several algorithmic changes have been introduced into G.729 which resulted in 50% drop in its complexity, enabling a DSP implementation with a complexity of about 10–12 MIPS. This paper describes the algorithmic changes which have been introduced in order to achieve the low complexity goal while meeting the terms of reference. Subjective tests have been performed by ITU-T in both the selection phase and the characterization phase and the results showed that the performance of G.729A is equivalent to both G.729 and G.726 at 32 kbit/s in most operating conditions; however, it is slightly worse in case of three tandems and in the presence of background noise. A breakdown of the complexities of both G.729 and G.729A is given at the end of the paper.

1. INTRODUCTION

There is currently a great interest in simultaneous transmission of voice and data in multimedia applications. At the request of Study Group 14 (SG 14) of ITU-T, an expert group (G.DSVD) was established in February 1995 within SG 15 for the specification of a new speech coding standard for use in digital simultaneous voice and data applications (DSVD). At that time, several standard codecs existed, or were being finalized in SG 15 (Recommendations G.728, G.723.1, and G.729). However, the modem experts in SG 14 felt that the complexity of these speech coding standards was prohibitive for integrating the modem algorithm and the speech coding algorithm on the same processor. This was reflected in the terms of reference for the new algorithm, where an upper limit of 10 MIPS was set on the complexity. It was also required that the RAM does not exceed 2K words and the ROM does not exceed 8K words. In terms of quality, it was required that the coder be as good as G.726 at 32 kbit/s in most operating conditions. The bit rate was not fixed but an upper limit of 11.4 kbit/s was set and preference was given to lower rates

In summer 1995, five contending codecs were submitted to the host lab for subjective testing (7.73 kbit/s with 15 ms speech frames from AT&T; 8.8 kbit/s with 10 ms frames from Audio Codes/DSP Group (AC/DSPG); 7.8 kbit/s with 15 ms frames from NTT; 8.0 kbit/s with 15 ms frames from Rockwell; and 8.0 kbit/s with 10 ms frames from the University of Sherbrooke (USH)). The contending codecs were tested in both North-Amerian English and Japanese languages (at COMSAT and NTT). The test results were discussed in the September 1995 meeting of G.DSVD where the codecs from AC/DSPG and USH came ahead of the other coders, and they were retained for further consideration. The coder from USH had the virtue of being bit-stream interoperable with G.729; i.e., speech encoded with G.729 can be decoded with the proposed DSVD coder, and vice versa. The interoperability with G.729 was considered important by SG 15, which felt that this will reduce the multiplicity of incompatible standards.

In the November 1995 meeting of SG 15, the coder from USH was finally selected, and the interoperabilty with G.729 had an important role in this decision. Instead of having a new Recommendation, it was decided to make the reduced complexity version of G.729 for DSVD an Annex A of G.729. It is now the standard speech codec in ITU-T V.70 series (DSVD).

Although the Annex A of G.729 was specifically recommended by the ITU-T for multimedia DSVD applications, the use of the codec is not limited to these applications. In fact, due to its interoperability with G.729, G.729A can replace G.729 in any of its applications when a complexity reduction is deemed necessary. Among the possible multimedia DSVD applications of G.729A are: multiparty multimedia conferencing, collaborative computing, telelearning and remote presentations, interactive games, file transfer during speech, mobile audiovisual services, telecommuting, teleshopping, and telemedicine. Another interesting potential application for G.729A is Internet telephony and Internet voice mail, where no standard speech coding algorithm exists. The relatively low complexity and low delay features of G.729A make it an attractive choice for such applications compared to G.723.1, the standard speech codec for GSTN visual telephony (H.324), which has at least twice the complexity and three times the delay.

In this paper, we describe the methods used to achieve the complexity reduction in the G.729 algorithm while maintaining a quality capable of meeting the terms of reference. Subjective test results from the selection phase as well as the characterization phase will be given. Finally, a breakdown of the codec complexity of both G.729 and G.729A will be given.

2. DESCRIPTION OF ALGORITHMIC CHANGES TO G.729

The general description of the coding/decoding algorithm of G.729A is similar to that of G.729 [1, 2, 3, 4]. The same conjugate-structure algebraic code-excited linear-predictive (CS-ACELP) coding concept is used. The coder operates on speech frames of 10 ms corresponding to 80 samples at a sampling rate of 8000 samples per second. For every 10 ms frame, the speech signal is analyzed to extract the parameters of the CELP model (linear-prediction filter coefficients, adaptive and fixed-codebook indices and gains). These parameters are encoded and transmitted. The bit allocation of the coder parameters is shown in Table 1. At the decoder, these parameters are used to retrieve the excitation and synthesis filter parameters. The speech is reconstructed by filtering this excitation through the short-term synthesis filter. After computing the reconstructed speech, it is further enhanced by a postfilter.

Parameter	*subfr 1*	*subfr 2*	*Total*
LSP coefficients			18
Pitch delay	8	5	13
Delay parity bit	1		1
Codebook positions index	13	13	26
Codebook signs index	4	4	8
Gains VQ	3+4	3+4	14
Total			80

Table 1. Bit allocation of the ITU-T 8 kb/s speech coder (G.729 & G.729A).

The LP analysis and quantization procedures as well as the joint quantization of the adaptive and fixed codebook gains are the same as G.729 [1, 2, 3]. The major algorithmic changes to G.729 are summarized below:

- The perceptual weighting filter uses the quantized LP filter parameters and it is given by $W(z) = \hat{A}(z)/\hat{A}(z/\gamma)$ with a fixed value of $\gamma = 0.75$.
- Open-loop pitch analysis is simplified by using decimation while computing the correlations of the weighted speech.
- Computation of the impulse response of the weighted synthesis filter $W(z)/\hat{A}(z)$, computation of the target signal, and updating the filter states are simplified since $W(z)/\hat{A}(z)$ is reduced to $1/\hat{A}(z/\gamma)$.
- The adaptive codebook search is simplified. The search maximizes the correlation between the past excitation and the backward filtered target signal (the energy of the filtered past excitation is not considered).
- The search of the fixed algebraic codebook is simplified.
- At the decoder, the harmonic postfilter is simplified by using only integer delays.

These changes are described in more detail in the following sections.

2.1. Perceptual weighting

Unlike G.729, the perceptual weighting filter is based on the quantized LP filter coefficients $\hat{a}_i$, and is given by

$$W(z) = \frac{\hat{A}(z)}{\hat{A}(z/\gamma)}, \qquad (1)$$

with $\gamma = 0.75$. This simplifies the combination of synthesis and weighting filters to $W(z)/\hat{A}(z) = 1/\hat{A}(z/\gamma)$, which reduces the number of filtering operations for computing the impulse response and the target signal and for updating the filter states. Note that the value of γ is fixed to 0.75 and the procedure for the adaptation of the factors of the perceptual weighting filter described in G.729 [4] is not used in G.729A.

The simplification of the weighting filter resulted in some quality degradation in case of input signals with flat response. In fact, the adaptation of the weighting factors was introduced in G.729 to improve the performance for such signals.

2.2. Open-loop pitch analysis

To reduce the complexity of the search for the best adaptive-codebook delay, the search range is limited around a candidate delay T_{ol}, obtained from an open-loop pitch analysis. This open-loop pitch analysis is done once per frame (10 ms). The open-loop pitch estimation uses the low-pass filtered weighted speech signal, $s_w(n)$, and is done as follows: in the first step, 3 maxima of the correlation

$$R(k) = \sum_{n=0}^{39} s_w(2n)s_w(2n-k) \qquad (2)$$

are found in the three ranges [20,39], [40,79], and [80,143]. The retained maxima $R(t_i)$, $i = 1, \ldots, 3$, are normalized through

$$R'(t_i) = \frac{R(t_i)}{\sqrt{\sum_{n=0}^{39} s_w^2(2n-t_i)}}, \quad i = 1, \ldots, 3. \qquad (3)$$

The winner among the three normalized correlations is selected by favoring the delays with the values in the lower range. This is done by augmenting the normalized correlations corresponding to the lower delay range if their delays are submultiples of the delays in the higher delay range.

Note that only the even samples are used in computing the correlations in Eq. (2). Further, in the third delay region [80,143] only the correlations at the even delays are computed in the first pass, then the delays at ±1 of the selected even delay are tested.

2.3. Closed-loop pitch search

The adaptive-codebook structure is the same as in G.729 [2, 5]. In the first subframe, a fractional pitch delay T_1 is used with a resolution of 1/3 in the range $[19\frac{1}{3}, 84\frac{2}{3}]$ and integers only in the range [85, 143]. For the second subframe, a delay T_2 with a resolution of 1/3 is always used in the range $[int(T_1) - 5\frac{2}{3}, int(T_1) + 4\frac{2}{3}]$, where $int(T_1)$ is the integer part of the fractional pitch delay T_1 of the first subframe.

Closed-loop pitch search is usually performed by maximizing the term

$$R(k) = \frac{\sum_{n=0}^{39} x(n)y_k(n)}{\sqrt{\sum_{n=0}^{39} y_k(n)y_k(n)}}, \qquad (4)$$

where $x(n)$ is the target signal and $y_k(n)$ is the past filtered excitation at delay k. In this reduced complexity version, the search is simplified by considering only the numerator in Eq. (4). That is, the term

$$R_N(k) = \sum_{n=0}^{39} x(n)y_k(n) = \sum_{n=0}^{39} x_b(n)u_k(n) \qquad (5)$$

is maximized, where $x_b(n)$ is the backward filtered target signal (correlation between $x(n)$ and the impulse response of the weighted synthesis filter $h(n)$) and $u_k(n)$ is the past excitation at delay k $(u(n-k))$.

For the determination of T_2, and T_1 if the optimum integer delay is less than 85, the fractions around the optimum integer delay have to be tested. The fractional pitch search is done by interpolating the past excitation at fractions $-\frac{1}{3}$, 0, and $\frac{1}{3}$, and selecting the fraction which maximizes the correlation in Eq. (5).

Simplifying the adaptive codebook search procedure resulted in some degradation compared to G729. The chosen pitch lag occasionaly differs by a fraction of 1/3 from that chosen in G.729.

2.4. Algebraic codebook: structure and search

The structure of the 17-bit fixed codebook is the same as G.729 [1, 2]. The algebraic codebook is a deterministic codebook whereby the excitation codevector is derived from the transmitted codebook index (no need for codebook storages).

The pulse amplitudes are preset using the same *signal-selected pulse amplitude* approach used in G.729. However, the pulse positions are determined using a new fast search procedure. In G.729, a fast search procedure based on a nested-loop search approach is used [1, 2, 5]. In that approach, only 1440 possible position combinations are tested in the worst case out of the 2^{13} position combinations (17.5%). In G.729A, in order to further simplify the search procedure, a smaller percentage of possible position combinations are tested using a depth-first tree search approach. In this approach only 320 position combinations are tested (3.9%).

About 50% of the complexity reduction in the coder part is attributed to the new algebraic codebook search (saving of about 5 MIPS). This was at the expense of slight degradation in the coder performance (about 0.2 dB drop in signal-to-noise ratio).

2.5. Post-processing

The post-processing is the same as in G.729 except for some simplifications in the adaptive postfilter.

The adaptive postfilter is the cascade of three filters: a long-term postfilter $H_p(z)$, a short-term postfilter $H_f(z)$, and a tilt compensation filter $H_t(z)$, followed by an adaptive gain control procedure [1, 4]. Several changes have been undertaken in order to reduce the complexity of the postfilter. The main difference from G.729 is that the long-term delay T is always an integer delay and it is computed by searching the range $[T_{cl}-3, T_{cl}+3]$, where T_{cl} is the integer part of the (transmitted) pitch delay in the current subframe.

The modifications in the postfiltering procedure resulted in a reduction of about 1 MIPS in the complexity.

3. CODER PERFORMANCE

The DSVD codec performance was determined in two phases. In the so-called *Selection Phase*, the five contenders were tested resulting in the selection of a single codec. This codec was then submitted to a more complete *Characterization Phase* subjective testing. In the Selection Phase, three experiments were performed on the contending codecs in both Japanese and North-American English languages, at NTT and COMSAT laboratories, respectively. Experiment 1 dealt with the characterization of the test codecs with input level variation and tandems (using flat speech). Experiment 2 characterized the codec performance for clear speech and in the presence of burst frame erasures (using modified IRS-weighted speech [6]). Experiment 3 dealt with the performance of the contending codecs in the presence of background noise (babble noise at 20 dB signal-to-noise ratio (SNR) and second talker at 15 dB SNR). In this article, only the results for USH codec are given for the English language [7]. Note that the tested USH coder is the same as the final version of G.729A exept for minor changes which were introduced to increase the common code between G.729 and G.729A. Details about test conditions and analysis can be found in [7].

Table 2 gives the subjective test results of Experiment 1 (modified IRS-weighted speech) of the Selection Phase for the English language [7]. An Absolute Category Rating (ACR) method was used [8]. The results are given in terms of *Mean Opinion Score* (MOS) and *equivalent Q* (Qeqv). The MNRU test conditions are used to derive a MOS versus Q curve from which the Qeqv value for each test condition is obtained. From the statistical analysis of the results, the USH codec met all the requirements, even the objective for the 3 tandem condition [7].

Coder	*Factor*	*MOS*	Qeqv
USH (8 kbit/s)	-16 dBov	3.61	26.53
USH (8 kbit/s)	-26 dBov	3.67	27.81
USH (8 kbit/s)	-36 dBov	3.52	24.92
USH (8 kbit/s)	2 tandems	3.13	20.60
USH (8 kbit/s)	3 tandems	2.51	16.08
G.726 (32 kbit/s)	-16 dBov	3.71	28.91
G.726 (32 kbit/s)	-26 dBov	3.59	26.07
G.726 (32 kbit/s)	-36 dBov	3.48	24.41
G.726 (32 kbit/s)	4 tandems	2.64	16.93
Source	none	3.990	•
MNRU	Q=30 dB	3.73	29.33
MNRU	Q=24 dB	3.49	24.51
MNRU	Q=18 dB	2.77	17.78
MNRU	Q=12 dB	1.91	12.13

Table 2. Test results of Experiment 1 of the selection phase for English language (performance in case of input level variations and tandems).

Table 3 gives the subjective test results of Experiment 2 (unweighted speech) of the Selection Phase for the English language [7] (an ACR method was used). From the statistical analysis of the results, the USH codec met the requirements for clear channel (equivalent to G.726) and for 3% frame erasure rate (less than 0.75 MOS degradation with respect to G.726). For the 5% FER, the codec was found statistically equivalent to 0.75 MOS degradation with respect to G.726.

Coder	*Factor*	*MOS*	Qeqv
USH (8 kbit/s)	0% FER	3.76	31.86
USH (8 kbit/s)	3% FER	3.18	26.61
USH (8 kbit/s)	5% FER	2.84	23.84
G.726 (32 kbit/s)	0% FER	3.65	30.74
Source	none	4.38	40.65
MNRU	Q30	3.59	30.21
MNRU	Q24	2.81	23.57
MNRU	Q18	2.20	18.52
MNRU	Q12	1.56	11.95

Table 3. Test results of Experiment 2 of the selection phase for English language (performance in clear conditions and burst frame erasures).

The subjective tests for the Characterization Phase of G.729A were performed in May 1996, for both Japanese and French languages at NTT and FT/CNET, respectively. The test consisted of three experiments [9]: Experiment 1 dealt with interworking between G.729 and G.729A (using an ACR method [8]); Experiment 2 dealt with the performance in the presence of background noise (using an comparison category rating (CCR) method [8]); and Experiment 3 dealt with the performance in the presence of channel errors and frame erasures (using an ACR method). Modified IRS weighted speech was used in all experiments. The results for the Japanese language are found in [10], where the conclusions of the three experiments are given below.

It was concluded from the results of Experiment 1 that [10]:

- No significant difference was found among the 4 possible interconnections of G.729/G.729A and the reference coder

(G.726 at 32 kbit/s).

- The scores for all eight combinations with 2-stage transcoding were higher than those for 4-stage transcoding of G.726 at 32 kbit/s.
- No significant difference was found between G.729A and G.726 at both high and low input levels.
- The quality of G.729A was slightly lower than that of G.729 under 3-stage transcoding.

It was concluded from the results of Experiment 2 that the scores of G.729A were slightly worse than those for G.729 and G.726 in both clear and background noise conditions, and that no siginificant differences were found for the possible combinations of the the two-stage transcoding of G.729 and G.729A under noise-free and background office noise conditions.

In Experiment 3, G.729A and G.729 were tested in case of 10^{-3} random bit errors and 3% and 5% random frame erasures in a quiet background, and also in babble background noise and office background noise conditions. In general, no difference was found betwen G.729A, G.729, and their interconnections.

4. CODEC IMPLEMENTATION AND COMPLEXITY

The reduced complexity CS-ACELP codec in G.729 Annex A specification consists of 16 bit fixed-point ANSI C code using the same set of fixed-point basic operators used to define G.729. A set of test vectors are provided as part of G.729A to insure that a certain DSP implementation is bit-exact with the fixed-point ANSI C code using basic operators. Basic operators are a C-language implementation of commonly found fixed-point Digital Signal Processor (DSP) assembly instructions. Describing an algorithm in terms of basic operators allows for easy mapping of the C-code to a certain DSP assembly language as well as for a rough estimate of the algorithmic complexity. A certain weight is associated with each basic operator which reflects the number of instruction cycles. Using these basic operators, the codec complexity was found to be 8.95 WMOPS (weighted million operations per second). A factor of 1.2–1.5 is usually used to estimate the complexity in MIPS (this depends on the DSP used and the actual function performed).

Both G.729A and G.729 were implemented on TI TMS320C50 DSP chip. In USH implementation, the full-duplex codec algorithm of G.729A required 12.4 MIPS while that of G.729 required 22.3 MIPS. The breakdown of the complexity of both G.729A and G.729 is given in Table 4, for both encoder and decoder. The complexity is given in terms of C50 MIPS and basic operator's WMOPS. In terms of memory occupation, G.729A required less than 2K RAM and 10K ROM while G.729 required about 2K RAM and 11K ROM.

5. CONCLUSION

This paper described the speech coding algorithm of Recommendation G.729 Annex A, which is the standard codec for multimedia digital simultaneous voice and data (DSVD). This algorithm is bit-stream interoperable with the algorithm specified in the main body of Recommendation G.729. It is an 8 kbit/s algorithm based on the CS-ACELP coding concept, and uses 10 ms speech frames. This algorithm resulted in about 50% drop in the complexity of G.729 at the expense of small degradation in the performance in case of three tandems and in the presence of background noise. Subjective test results performed in the standard's Characterization Phase showed that there is no difference among G.726 at 32 kbit/s and the four possible combinations of G.729/G.729A. The codec was implemented on a TI TMS320C50 fixed-point DSP chip where a complexity of 12 MIPS was required (full-duplex), with less than 2 K RAM and less than 10 K ROM. A full-duplex implementation of G.729 requited 22 MIPS. A complexity down to 10 MIPS can be easily obtained using a more recent chips such as TMS320C540.

	WMOPS		C50 MIPS	
Function	*G.729*	*G.729A*	*G.729*	*G.729A*
Pre-processing	0.20	0.20	0.226	0.226
LP analysis & quant.	2.88	2.35	3.808	3.259
Pitch analysis	4.28	2.37	5.016	2.732
Algebraic codebook	6.35	1.86	8.406	3.046
Gains VQ	0.46	0.46	0.643	0.643
Other	0.21	0.08	0.278	0.112
Total (coder)	14.38	7.32	18.377	10.019
Decoder	0.68	0.68	1.133	1.133
Postfilter	2.13	0.73	2.539	1.000
Post-processing	0.22	0.22	0.266	0.266
Total (decoder)	3.03	1.63	3.938	2.399
Total (duplex)	17.41	8.95	22.315	12.418

Table 4. Breakdown of the codec complexity (worst case) for G.729 and G.729A in terms WMOPS and TMS320C50 MIPS.

REFERENCES

[1] Draft Recommendation G.729, "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic Code-Excited Linear-Prediction (CS-ACELP)," International Telecommunication Union, Telecommunications Standardization Sector, 1995.

[2] R. Salami et al., "Description of the proposed ITU-T 8 kb/s speech coding standard," *Proceedings IEEE Speech Coding Workshop*, Annapolis, Sept. 1995, pp. 3–4.

[3] A. Kataoka et al., "LSP and gain quantization for the proposed ITU-T 8 kb/s speech coding standard," *IEEE Speech Coding Workshop*, Annapolis, Sept. 1995, pp. 7–8.

[4] D. Massaloux and S. Proust, "Spectral shaping in the proposed ITU-T 8 kb/s speech coding standard," *IEEE Speech Coding Workshop*, Annapolis, Sept. 1995, pp. 9–10.

[5] R. Salami, C. Laflamme, J-P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 808–816, Aug. 1994.

[6] ITU-T Recommendation P.48, "Specification for an intermediate reference system," volume V of Blue Book, pp. 81-86, ITU, Geneva 1989.

[7] ITU-T SG 15 contribution, "Final test report of DSVD Experiments 1, 2 and 3 for North-American English," Source: COMSAT, Geneva, November 1995.

[8] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," Geneva, May 1996.

[9] SQEG contribution SQ-35.96, "Subjective test plan for characterization of an 8 kbit/s speech codec for DSVD applications," ITU-T SG 12, March 1996.

[10] ITU-T SG 15 contribution, "Results of characterization testing using Japanese language for draft Annex A to Recommendation G.729 (Low-complexity CS-ACELP for DSVD applications)," Source: NTT, May 1996.