16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

ON THE ITU-T G.729.1 SILENCE COMPRESSION SCHEME

Panji Setiawan(1), *Stefan Schandl*(2), *Hervé Taddei*(3)*, *Hualin Wan*(3), *Jinliang Dai*(3),
Libin Zhang(3), *Deming Zhang*(3), *Jun Zhang*(3), *and Eyal Shlomot*(4)

(1) Siemens Enterprise Communications GmbH & Co. KG; Hofmannstr. 51, 81379, Munich, Germany
(2) Siemens AG, Austria; Erdberger Lände 25, A-1031, Vienna, Austria
(3) Huawei Technologies Co., Ltd.; Bantian, Longgang District, 518129, Shenzhen, P.R. China
(4) Comango Technologies LLC; 216 Quincy Ave, Long Beach, CA, USA

ABSTRACT

Silence compression scheme is essential for efficient voice communication systems. It allows a significant reduction of transmission bandwidth during silence period where only parametric descriptions of the background noise are transmitted. A silence compression scheme includes a voice activity detection (VAD), a discontinuous transmission (DTX), a silence insertion descriptor (SID), and a comfort noise generator (CNG) module. In this paper, we describe a silence compression scheme for ITU-T Recommendation G.729.1 which employs a unique scalable SID frame structure. This scalable frame structure consists of a lower band core layer, a lower band enhancement layer, and a higher band layer. The scheme is optimized for ITU-T G.729.1 and is interoperable with ITU-T G.729 Annex B.

1. INTRODUCTION

A silence compression scheme for the ITU-T G.729.1 Recommendation [1] is described in this paper. Current multimedia and personal communication services are widely deployed with silence compression schemes, which allow the reduction of transmission bandwidth by taking advantage of the inactive periods of speech. Such schemes generally use a voice activity detection (VAD) module to distinguish between the active and inactive speech period. During the inactive speech periods, a lower bitrate is achieved by stopping the transmission and only sending silence insertion description (SID) updates when changes in the background noise characteristics are detected. The SID frames contain a representation of the background noise characteristics which are used by the decoder to generate a background noise with similar characteristics.

The SID frame structure of our silence compression scheme consists of 3 layers, a lower band core layer compatible with G.729 Annex B [2, 3], a lower band enhancement layer, and a higher band layer.

The paper is organized as follows: Section 2 describes the standardization process of G.729 and G.729.1 codec including the corresponding silence compression scheme. Section 3 gives an overview of the silence compression scheme. Section 4 presents a detailed description of the DTX module followed by the description of the SID bitstream structure in Section 5 and of the CNG in Section 6. A method of handling bitrate switching during the DTX operation, as an extension of the bitrate switching handling of G.729.1 codec, is presented in Section 7. The performance evaluation of the proposed silence compression scheme is outlined in Section 8 and Section 9 concludes our paper.

2. THE ITU-T G.729 AND G.729.1 STANDARDS

ITU-T G.729.1 codec (standardized in May 2006) is an 8-32 kbit/s scalable wideband (50-7000 Hz) extension of the G.729 codec [4]. The scalable structure of its bitstream allows a flexible bitrate and bandwidth adjustment during the transmission. By default, the encoder input and decoder output are sampled at 16000 Hz.

The codec employs a three-stage structure: an embedded code-excited linear-prediction (CELP) codec, a time-domain bandwidth extension (TDBWE), and a predictive transform coding, which is referred to as the time-domain aliasing cancellation (TDAC). ITU-T G.729.1 codec uses 20 *ms* superframes, but its embedded CELP stage uses 10 *ms* frames (as in G.729), which means that two 10 *ms* CELP frames are processed per 20 *ms* superframe.

The bitstream produced by the encoder consists of 12 embedded layers. The first layer is the core layer with a bitrate of 8 kbit/s. This layer is compatible with G.729 bitstream. The second layer is a narrowband enhancement layer, which adds 4 kbit/s to reach the rate of 12 kbit/s. The other 10 layers are the wideband enhancement layers, each adds 2 kbit/s. A high level description of G.729.1 can be found in [5].

To help introducing wideband communications, G.729.1 has been built with a core interoperable with G.729, a widely deployed narrowband speech codec [4] for voice over IP (VoIP). However, while G.729 includes a silence compression scheme in G.729 Annex B [2], G.729.1 was initially standardized without a silence compression scheme.

Following the standardization of G.729.1, a new work item was launched in ITU-T Study Group 16 Question 10 (Q10/16) to extend G.729.1 with a silence compression scheme for VoIP applications such as enterprise networks. A term of reference (ToR) [6] was established, with the VAD of G.722.2 [7] as a provisional VAD for the test. While several companies indicated their intentions to participate in a qualification phase, only Huawei Technologies and Siemens Enterprise Communications confirmed their intentions to participate and these two companies decided to collaborate on this work item. Therefore, Q10/16 decided to skip the qualification phase and to launch an optimization/characterization phase. The characterization phase quality assessment test plan was drafted by Q7/12 speech quality experts group (SQEG) and the listening test phase took place in March 2008. The test results were then submitted to Q7/12 for a

(∗) H. Taddei was with Nokia Siemens Networks when this work was done.

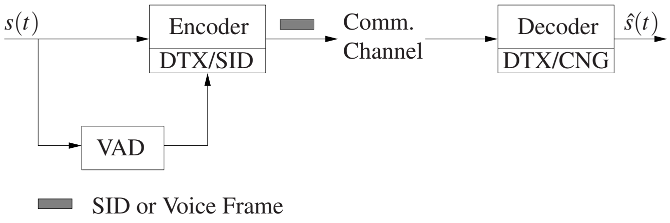

Figure 1: Typical structure of a silence compression scheme, $s(t)$ and $\hat{s}(t)$ denote the input and output signal, respectively.

thorough analysis and finally recommended Q10/16 to go forward with the standardization of this scheme. A draft request for comments (RFC) [8] has been compiled by the Internet engineering task force (IETF) to define the bitstream of silence compression scheme in the real-time transport protocol (RTP) format for G.729.1.

3. SILENCE COMPRESSION SCHEME GENERAL DESCRIPTION

A silence compression scheme typically consists of several modules, i.e., VAD, DTX, SID, and CNG modules as shown in Figure 1. In the development work of the G.729.1 silence compression scheme, the input VAD decision was obtained from the VAD module of the standardized adaptive multi rate - wideband (AMR-WB) codec [7]. In general, the VAD module can be chosen from many other existing VAD modules.

At the transition from active to inactive speech period, the encoder features a hangover phase, during which the onset of inactive superframe is still coded by voice or active speech parameters and at the same time the parameters for the inactive speech period are being learned (noise learning phase). The first step in reducing the source bitrate is already accomplished in that only the layers up to 14 kbit/s are encoded during this hangover phase. At the end of the hangover phase an initial SID frame, containing the information about the background noise estimated during the hangover phase, is transmitted. The hangover phase takes a length of five superframes. Further transmission of SID frames is decided by the DTX handler, which is described in Section 4.

The generation of background noise during the inactive speech period is done by the CNG module. The CNG uses the information from the SID frames to generate the comfort noise. It will retain the information from the last SID frame and continue generating the noise using this information as long as there is no SID frame update.

4. DISCONTINUOUS TRANSMISSION (DTX)

The DTX handler is designed to trigger the transmission of an SID frame if the characteristics of the background noise change significantly. In addition to that, self-triggered transmission of an SID frame can be forced such that a minimum frame rate can be guaranteed. This feature has been designed for applications that might need a minimum payload (alive indication). The minimum SID frame rate can be supplied as a codec (encoder) parameter.

As the structure of the codec for the inactive speech period is similar to that of the active speech case at up to 14 kbit/s, so is the DTX handler structured into narrowband and wideband parts.

4.1 Lower Band DTX

For each inactive 20 *ms* superframe, the lower band DTX module indicates the need of sending an update of inactive parameters, by measuring the perceptual changes during the inactive speech period. For each of the two 10 *ms* frames of each 20 *ms* superframe of an inactive speech period, the algorithm is based on the same approach as of the DTX in G.729 Annex B, in comparing the linear prediction coding (LPC) filter and the energy of the current frame to the previously transmitted LPC filter and energy. The Itakura distance measure is used for the comparison of the LPC filters and the absolute distance in the dB domain is used for the comparison of the energies, similar to G.729 Annex B. However, very minor modifications of the LPC filter and the energy estimate are introduced, as well as minor adjustments of the thresholds used in the decision. The lower band DTX flag for the entire 20 *ms* superframe, NB_flag_change is set if perceptual significant change is detected in either the first 10 *ms* frame or the second 10 *ms* frame.

4.2 Higher Band DTX

The wideband extension layer for the inactive speech period is very similar to the TDBWE algorithm used in the 14 kbit/s layer of G.729.1. As described in [9] the TDBWE algorithm uses a time envelope $T^m_{env}(i)$ and a frequency envelope $F^m_{env}(j)$ to describe the properties of the signal in time and frequency domains where m denotes the superframe index. Since only slow temporal variations of energy can be described by the SID scheme, the values of the time envelope are replaced by their mean value w.r.t. a frame's length:

$$\overline{T}^m_{env}(i) = \overline{T}^m = \frac{1}{16}\sum_{i=0}^{15} T^m_{env}(i), \qquad (1)$$

in case of an inactive superframe. These parameters, $\overline{T}^m$ and $F^m_{env}(i)$, are lowpass-filtered according to

$$\tilde{T}^m(i) = \alpha_{tenv} \cdot \overline{T}^m(i) + (1 - \alpha_{tenv}) \cdot \tilde{T}^{m-1}(i), \qquad (2)$$

and

$$\tilde{F}^m_{env}(j) = \alpha_{fenv} \cdot F^m_{env}(j) + (1 - \alpha_{fenv}) \cdot \tilde{F}^{m-1}_{env}(j), \qquad (3)$$

where $\alpha_{fenv} = 0.25$ and $\tilde{T}^m(i)$ and $\tilde{F}^m_{env}(j)$ denote the filtered time and frequency envelope parameters, respectively. The initial values $\tilde{T}^{m-1}(i)$ and $\tilde{F}^{m-1}_{env}(j)$ are derived during the noise learning phase by averaging the envelope parameters $\overline{T}^m(i)$ and $F^m_{env}(j)$ over a couple of superframes. During the active speech period, a mean time envelope value M_T, and mean-removed time and frequency envelope parameter sets, $T^M_{env}(i)$ and $F^M_{env}(j)$ will be used in the TDBWE quantization [9]. During the inactive speech period, the corresponding modified parameter sets $\tilde{M}_T$, $\tilde{T}^M_{env}(i)$, and $\tilde{F}^M_{env}(j)$ are calculated and quantized.

The variations of the filtered parameters are monitored throughout the inactive speech period. Two kinds of variations are tracked, namely the short-term changes in order to have an indication of transitions in the properties of the background noise and the long-term changes in order to keep track of a temporal drift in the same properties. If any of the parameters monitored exceeds an associated threshold the flag WB_flag_change is set to 1.

NB_flag_change	WB_flag_change	$flag\_change$
0	0	0
1	0	combined test
0	1	combined test
1	1	1

Table 1: Combined DTX decision flag.

4.3 Combined DTX Decision

The final DTX decision flag, $flag\_change$ is determined based on both NB_flag_change and WB_flag_change, using the decision table shown in Table 1. When the NB_flag_change and the WB_flag_change do not agree, the *combined test* is used to avoid too frequent SID updates which can unnecessarily reduce the bandwidth saving. A decision measure d is calculated according to:

$$\begin{aligned} d &= w_1 \cdot \left|T^{sid} - \tilde{T}^m(i)\right| + w_2 \cdot \sum_{j=0}^{11} \left|F^{sid}(j) - \tilde{F}^m_{env}(j)\right| + \\ &+ w_3 \cdot C_f + w_4 \cdot C_g, \end{aligned} \tag{4}$$

where $w_1, \cdots, w_4$ are are optimally tuned weights. For the upper band, T^{sid} is the time envelope of the last SID superframe, $F^{sid}(j)$ is the j^{th} frequency envelope of the last SID superframe. For the lower band, C_f is the ratio of the Itakura distance between the previous SID LPC filter, current LPC filter and the threshold, which is expressed by:

$$C_f = \frac{\sum_{j=0}^{N} R_a(j) \times R^t(j)}{E_t \times thr_1}, \tag{5}$$

(see Eq. B.12 in [2]) where $R_a(j)$ is a function derived from the autocorrelation of the coefficients of the SID filter and $R^t(j)$ denotes the cumulative sum of the autocorrelation functions. C_g is calculated as follows:

$$C_g = \frac{|E_t - E_q^{sid}|}{thr_2}, \tag{6}$$

(see Section B.4.1.4 in [2]) where E_q^{sid} denotes the previously decoded SID log-energy and E_t is the current superframe residual energy. When $d < 1$, $flag\_change$ is set to 0, otherwise $flag\_change$ is set to 1.

5. SILENCE INSERTION DESCRIPTOR (SID) STRUCTURE

The SID structure consists of three embedded layers, i.e., a lower band core layer, a lower band enhancement layer, and a higher band layer, having 15, 9, and 19 bits, respectively.

5.1 Lower Band Core Layer

The structure of the core SID is identical to the SID of G.729 Annex B, which allows G.729.1 to decode SID frames generated by G.729 Annex B and for G.729 Annex B to decode SID core frames generated by G.729.1. This SID structure uses 10 bits in split vector quantizer (VQ) to describe the LPC filter and 5 bit scalar quantizer for the energy.

5.2 Lower Band Enhancement Layer

The LPC filter representation is improved by allocating an additional 6 bits to the line spectral frequency (LSF) quantization. The third stage codebook is constructed from the second stage codebook of G.729 using an index-mapping approach as in G.729 Annex B. The vectors in this third stage codebook are further multiplied by a gain factor, obtained from an 8-entries scalar codebook, which is indexed by the upper 3 bits of the energy index of the core layer energy quantizer. The enhancement layer codebook quantizes the vector of LSF quantization errors, obtained by subtracting the core layer quantized LSF vector from the original LSF vector. Another 3 bits are used to improve the energy scalar quantizer, where 8 quantization levels are uniformly added between each of the quantization level of the core layer. Below -4 dB (the lowest quantization level of the core layer) 8 quantization levels spaced at 1 dB are added down to -12 dB.

5.3 Higher Band Layer

The encoding of the higher band part [4-7 kHz] of the signal during inactive period is very close to the TDBWE module used in [1]. The main difference between voice and silence coding is due to the nature of the signal, notably very slow temporal variation of parameters or infrequent update of the parameters, which results in the time envelope considered as being constant in time. The mean removed time envelope parameter vectors, consequently, are set to zero, $\tilde{T}^M_{env}(i) = 0$, and the bits that were allocated to the time envelope need not to be spent on the inactive superframe case.

6. COMFORT NOISE GENERATOR (CNG)

6.1 Lower Band CNG

The lower band CNG algorithm is very similar to the algorithm used in G.729 Annex B. However, several changes were introduced for the reconstruction of the parameters, the parameter interpolation, the control of the energy, and the smoothing of the transitions between the active signal and CNG.

For the first inactive superframe, the energy and the spectral parameters for the first frame within that superframe are estimated from the parameters of the hangover period. The energy and the spectral parameters extracted from the first SID information are extrapolated and used for the second frame of that superframe and for the next non-transmitted superframes. Similar extrapolation is performed each time a new SID frame is received.

The extrapolation uses d_{sid}, the distance (in frame numbers) between the currently received SID and the previously received SID. The energy is extrapolated by

$$E_{cur} = E_{pre} + \frac{E_{sid} - E_{pre}}{|k - d_{sid}| + 1}, \tag{7}$$

where E_{cur} is the energy of the current (k^{th}) frame, E_{sid} is the energy delivered by the last SID, and E_{pre} is the energy delivered by the SID before the last one. The spectral parameters are interpolated in a similar way, but with an added dither, according to the following procedure. First, a dither parameter for each LSF index i is calculated by:

$$D_{lsf}(i) = \frac{\omega_{sid}(i) - \omega_{pre}(i)}{|k - d_{sid}| + 1}, \tag{8}$$

where $\omega_{sid}(i)$ is the i^{th} LSF received from the last SID and $\omega_{pre}(i)$ is the i^{th} LSF received by the SID before the last one. Then the current frame i^{th} LSF is calculated by:

$$\omega_{cur}(i) = \omega_{pre}(i) + D_{lsf}(i) + rand\left(-\frac{D_{lsf}(i)}{2}, \frac{D_{lsf}(i)}{2}\right),$$

where $rand(a,b)$ returns a random number between a and b.

The excitation at the transition between the last active superframe and the first CNG superframe is smoothed by an overlap-and-add approach between an excitation superframe generated using parameters from the last active superframe and an excitation generated using the noise parameters.

The excitation will then be shaped at every frame by a shaping filter to improve the perceptual quality of the synthesized signals. The excitation is generated as in G.729 Annex B and is filtered by the reconstructed LPC filter to create the comfort noise. The energy of the generated comfort noise is slightly attenuated for a long period of distinctly stationary noise.

6.2 Higher Band CNG

The higher band extension layer for the inactive speech period is very closely matched to the active speech [9], but there is an important difference in the generation of the excitation signal. The TDBWE algorithm for the active frame derives the parameters voiced gain g_v and unvoiced gain g_{uv} satisfying

$$g_v^2 + g_{uv}^2 = 1, \tag{9}$$

from the core layer CELP codec's pitch parameter as well as the fixed and adaptive codebook gains. These parameters are used to synthesize an artificial excitation signal, which is mixed from a voiced and an unvoiced contribution. For the inactive speech case there is no such voiced contribution to the excitation signal since the CELP layer does not provide the pitch parameters and adaptive codebook (voiced) contribution. Hence, the gain parameters of the TDBWE are set to $g_v = 0$ and $g_{uv} = 1$ for the inactive speech case. As a consequence, the generation of the voiced contribution to the excitation signal as described in [9] is omitted and the white noise signal remains as the only source of excitation.

7. BITRATE SWITCHING

The bitrate switching method from G.729.1 has been extended to the DTX/CNG operation mode. It uses a mechanism to gather information on the wideband speech presence. If wideband speech is mostly observed during the conversation, the switching is directed to the wideband. The same method applies for the narrowband case. Switching from narrowband to wideband requires a shorter transition period (100 *ms*) compared to the transition in the active speech period. Unlike the immediate transition from wideband to narrowband condition as applied during the active speech period, during the DTX/CNG mode it requires 1 *s* transition period by expanding the bandwidth and creates new spectral components to high frequencies based on the last received TDBWE parameters, then fades out high frequencies using time varying parameters.

8. PERFORMANCE EVALUATION

8.1 Complexity and Memory Requirements

The proposed silence compression scheme adds around 0.27 WMOPS computational complexity to the G.729.1 during *active* transmission. The active transmission is defined when the G.729.1 encoded frames are being transmitted. This calculation is evaluating the additional complexity which occurs during the hangover period. During *inactive* transmission, i.e., when the SID frames are being transmitted and the CNG is active, the complexity is shown around 16 WMOPS, which is less than half the G.729.1 complexity (around 35 WMOPS). The complexity is measured at 32 kbit/s.

In addition to that, the additional RAM requirements are 0.34 and 0.03 *kwords* for the static and dynamic RAM, respectively. The additional data ROM is 0.287 *kwords* and the program ROM is given as 9557 which indicates an increase of 1232 *basic operations* and *function calls*. Note that the above requirements do not include the VAD module.

8.2 Listening Test Setup

Two listening laboratories were involved in testing the silence compression scheme. Beijing Institute of Technology (BIT) conducted the test in Chinese language and France Telecom (FT) in French language. In each laboratory, two listening tests were conducted, one in narrowband and one in wideband at three different bitrates, i.e., 12, 22, and 32 kbit/s. A total of 32 native listeners divided into 4 groups of 8 listeners, were required for each test.

Each laboratory provided the speech material which consisted of 16 kHz speech samples obtained from 3 male and 3 female speakers. The listening test and the processing plan can be found in [10, 11]. For filtering and downsampling, the processing functions from the software tool library (STL) [12] were used.

Noisy speech conditions were obtained by adding noise to the clean speech signal as specified in the processing plan. Three different noises were required, i.e., office noise, babble noise with 40 voices, and babble noise with 128 voices. The resulting noisy speech signal was adjusted to 20, 30, and 20 dB signal to noise ratio (SNR) level, respectively.

8.3 Test Results and Analysis

The listening tests have been successfully conducted in March 2008 and the results have been analyzed by Q7/12 in April 2008. The proposed silence compression scheme has passed all quality requirements and some quality objectives verification as shown in Tables 2 and 3, respectively. The chosen procedure to assess the scheme is the degradation category rating (DCR) as described in [13]. The quantity evaluated from the scores is represented by the symbol DMOS (degradation mean opinion score).

The quality requirements were tested using the *Poor or Worse* (PoW) procedure. This procedure compares the amount of *annoying* and *very annoying* votes between the reference G.729.1 codec (DTX OFF) and the proposed scheme (DTX ON). Having PoW votes of 0, 1, 2, and 3 for the reference will set the maximum allowed PoW votes for the proposed scheme to 19.2, 20.2, 21.2, and 22.2, respectively. The results in Table 2 show much lower PoW votes for the proposed scheme than the maximum allowed PoW votes.

# of PoW Votes		Office, 20 dB		Babble 40, 30 dB		Babble 128, 20 dB	
		FT	BIT	FT	BIT	FT	BIT
12 kbit/s	DTX OFF	-	-	2	1	1	1
	DTX ON	-	-	2	-	-	-
22 kbit/s	DTX OFF	1	-	2	1	3	2
	DTX ON	3	2	1	1	2	7
32 kbit/s	DTX OFF	1	3	-	-	2	1
	DTX ON	1	1	-	-	-	3

Table 2: Quality requirement test results using the *Poor or Worse* (PoW) procedure.

DMOS		Clean Speech		Office, 20 dB		Babble 40, 30 dB		Babble 128, 20 dB	
		FT	BIT	FT	BIT	FT	BIT	FT	BIT
12 kbit/s	DTX OFF	4.78(0.08)	4.78(0.08)	4.46(0.09)	4.66(0.01)	4.66(0.08)	4.72(0.09)	4.52(0.11)	4.59(0.10)
	DTX ON	4.71	4.74	4.48	4.72	4.71	4.70	4.45	4.63
22 kbit/s	DTX OFF	4.57(0.07)	4.54(0.10)	4.42(0.09)	4.54(0.10)	4.49(0.08)	4.54(0.09)	4.35(0.08)	4.49(0.12)
	DTX ON	4.59	4.63	4.29	4.38	4.42	4.44	4.45	4.30
32 kbit/s	DTX OFF	4.72(0.06)	4.67(0.09)	4.59(0.08)	4.59(0.10)	4.62(0.07)	4.68(0.09)	4.57(0.09)	4.60(0.10)
	DTX ON	4.67	4.69	4.49	4.47	4.54	4.59	4.52	4.43

Table 3: Quality objective test results using the *not worse than* procedure.

The quality objectives were evaluated using the *not worse than* procedure. This procedure uses a one-sided t-test to set the lower limit of the proposed scheme DMOS value. The lower limit is obtained from Table 3 by subtracting the value in the brackets from its reference DMOS value. As shown in the table, all narrowband tests and some of the wideband tests pass this quality verification procedure. The *worst* wideband test results are showing $\Delta_{DMOS} = 0.04$ and $\Delta_{DMOS} = 0.07$ from the lower limit in FT and BIT tests, respectively. Note that the results also tabulate the performance in clean speech.

Both quality verification procedures confirm the excellent performance of the proposed silence compression scheme. In addition to that, the evaluation gathered on the test materials at 32 kbit/s indicates an average bitrate at around 19 kbit/s was obtained using the proposed scheme with the percentage of speech activity at around 56%.

9. CONCLUSION

In this paper we have described the latest silence compression scheme standardized in the ITU-T extending the G.729.1 functionalities. It was shown that the proposed silence compression scheme achieved a high quality performance. It features an embedded SID structure and offers a full interoperability operation with the G.729 Annex B silence compression scheme and supports all G.729.1 bitrate modes.

10. ACKNOWLEDGMENTS

The authors would like to thank Claude Lamblin and Catherine Quinquis for their support during the development work and also Alan Sharpley (Dynastat) and BIT especially for providing the noise files used in the listening tests.

REFERENCES

[1] ITU-T Rec. G.729.1, *G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729*. 2006.

[2] ITU-T Rec. G.729 Annex B, *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP); A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70*. 1996.

[3] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, “ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications,” *IEEE Communications Magazine*, vol. 35, pp. 64–73, Sept. 1997.

[4] ITU-T Rec. G.729, *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*. 1996.

[5] 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. ICASSP 2007*, Honolulu, Hawaii, USA, April 15-20, 2007, pp. IV-529–IV-532.

[6] ITU-T Q10/16 Doc. TD 297 R1 (WP 3/16) Annex Q10.F, *Terms of Reference (ToR) and Time schedule for ITU-T G.729.1 DTX/CNG scheme*. 2008.

[7] ITU-T Rec. G.722.2, *Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*. 2003.

[8] IETF Internet-Draft draft-ietf-avt-rfc4749-dtx-update-00, *G.729.1 RTP payload format update: DTX support*. 2008.

[9] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaume, and S. Ragot, “Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2496–2509, Nov. 2007.

[10] ITU-T Q7/12 Doc. TD AH-08-17, *G.729.1 DTX/CNG extension characterization Quality Assessment Test Plan*. 2008.

[11] ITU-T Q10/16 Doc. AC-0801-Q10-38R1, *Processing Test Plan for the ITU-T G.729.1 DTX/CNG scheme optimization/characterization phase*. 2008.

[12] ITU-T Rec. G.191, *Software tools for speech and audio coding standardization*. 2005.

[13] ITU-T Rec. P.800, *Methods for subjective determination of transmission quality*. 1996.